TESTING MULTISPECIES COALESCENT SIMULATORS WITH SUMMARY STATISTICS

By

Hector Daniel Baños Cervantes.

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Statistics

University of Alaska Fairbanks

December 2018

APPROVED:

Elizabeth Allman, Committee Co-Chair
John Rhodes, Committee Co-Chair
Scott Goddard, Committee Member
Julie McIntyre, Committee Member
Ron Barry, Committee Member
Anthony Rickard, Department Chair
Department of Mathematics and Statistics

Abstract

The Multispecies coalescent model (MSC) is increasingly used in phylogenetics to describe the formation of gene trees (depicting the direct ancestral relationships of sampled lineages) within species trees (depicting the branching of species from their common ancestor). A number of MSC simulators have been implemented, and these are often used to test inference methods built on the model. However, it is not clear from the literature that these simulators are always adequately tested. In this project, we formulated tools for testing these simulators and use them to show that of four well-known coalescent simulators, Mesquite, Hybrid-Lambda, SimPhy, and Phybase, only SimPhy performs correctly according to these tests.

Acknowledgments

I would like to express the deepest appreciation to my advisers Dr. John A. Rhodes and Dr. Elizabeth Allman for letting me work with them, for their patience and for contributing to this work.

I would also like to express appreciation to my committee, Dr. Scott Goddard, Dr. Ron Barry, and Dr. Julie McIntyre, for taking time to answer questions, to review and to give me suggestions for this work.

In addition, I would like to thank Dr. Margaret Short and Dr. Ed Bueler for taking time to answer questions and help me. Also, I would like to thank Tony Knowles for carefully reading and suggesting revisions.

Contents

List of Figures

List of Tables

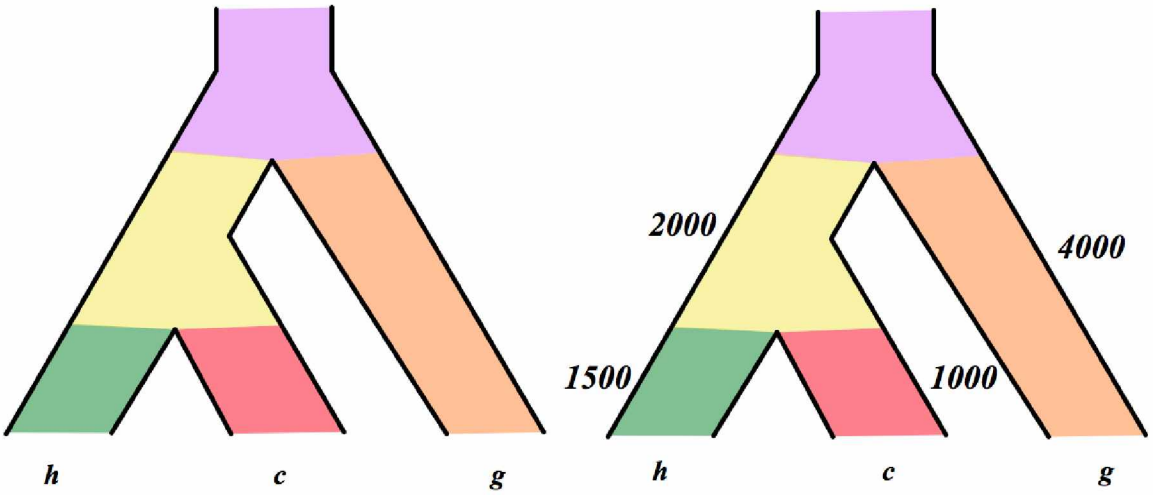

FIGURE 1. (Left) A species tree relating human (h), chimp (c), and gorilla(g) with Newick notation $((h, c), g)$. (Right) A metric species tree version of the tree on the left with Newick notation $((h{:}1000, c{:}1500){:}2000, g{:}4000)$.

1. INTRODUCTION

An important goal of phylogenetics is to infer evolutionary relationships between species. One of the main tools used to depict such relationships is a *species tree*. A species tree S is a branching diagram composed of "tubes", often called edges, representing populations. These tubes are merged together to depict speciation events. Figure 1 depicts a species tree, where the *leaves* (the bottom of the populations labelled as h, c, and g) represent the species to be related, and the rest of the populations represent ancestors of the leaves. The *root* (the population on the top of the diagram) is the most recent common ancestor of all the species that are being related. Often species trees are denoted using Newick notation. For example, the tree shown on the left of Figure 1 has Newick notation $((h, c), g)$, see [AR05] for more details on Newick notation.

Phylogenetics not only studies how species are related, but also how distant these relationships are. We can assign to any tree S a function $\tau : E(S) \to [0, \infty)$, where $E(S)$ is the set of edges (i.e populations) on S, that encodes the number of generations in a population. This provides a sense of *edge length*, or how much time, in number of generations, has passed between speciation events. We refer to the pair (S, τ) as a *metric species tree*. We can extend the Newick notation to include this metric information. For example, the metric Newick notation of the tree on the right of Figure 1 is $((h{:}1500, c{:}1000){:}2000, g{:}4000)$.

For any metric species tree (S, τ), one can also assign to each population e a function $N_e(t) : [0, \tau(e)) \to \mathbb{R}_{>0}$, where $\tau(e)$ is the edge length of e. The function $N_e(t)$ represents the population size in edge e at time t, where time is measured in number of generations backwards in real time (towards the top of the diagram). One also assigns a population function $N_r(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ to the root of the tree, representing the size of the population at the root and above it at time t, where time is also measured in number of generations.

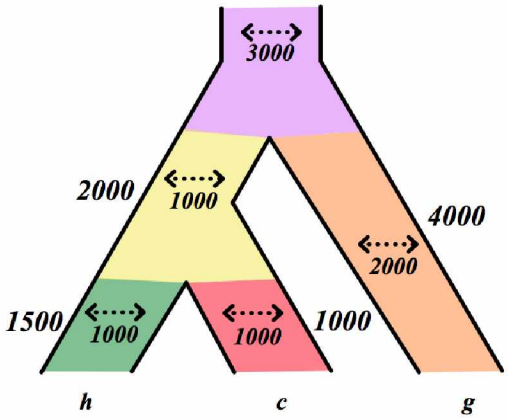

FIGURE 2. The species tree ((*h*:1000#1000, *c*:1500#1000):2000#1000, *g*:4000#2000)#3000, with length of edges in generations and population sizes in individuals.

For technical reasons $N_e(t)$ must be bounded above ($N_e(t) \leq B$ for all $e \in E(S)$ and some $B < \infty$), and $1/N_e(t)$ must be integrable over finite intervals.

If in a given tree these population size functions on edges are constant, one can extend the metric Newick notation to include the population size by adding a pound sign (#) to the respective edge. For example,

$$((h{:}1500\#1000, c{:}1000\#1000){:}2000\#1000, g{:}4000\#2000)\#3000 \tag{1}$$

corresponds to the species tree on the right of Figure 1 where the population size above the root is 3000 and all edges have population size of 1000 except the edge from the root to g which has a population of 2000 individuals. Note that we can depict a species tree using the length of a tube to represents to the number of generations, and the width to represent population size. Figure 2 depicts the tree in (1) not drawn to scale.

There are different ways to infer metric species trees. Before DNA sequencing was available, biologists used morphological data to determine species trees. One problem with this approach is *convergent evolution*, a process where organisms not necessarily closely related evolve similar traits as a result of adaption in similar environments (see for example [Ree+10]). Convergent evolution is less likely to affect inference methods that use genomic data, so generally methods based on DNA or protein sequences are now preferred; see for example [SLA16; CKC07; DS05]. One common approach for inference methods is to first construct *gene trees*, which are trees showing the relation between genes of different species using DNA to trace ancestry, and then use gene trees to infer species trees. A gene tree, denoted by (G, t), is often depicted using a metric tree diagram, where G contains the "topological" information (the shape of the tree) and t the metric information of G. These trees are drawn using lines, which represent direct parental ancestry of the sampled genes. In the right of Figure 3 we see a metric gene tree relating genes H, G and C.

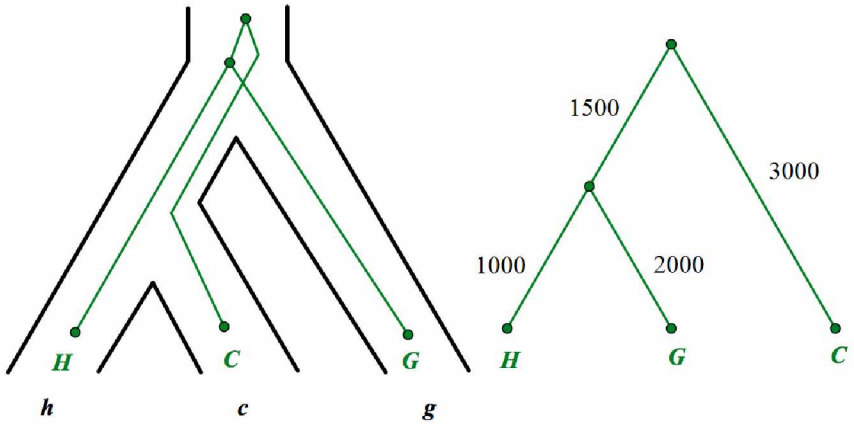

FIGURE 3. (Left) A gene tree $((H, G), C)$ within the species tree $((h, c), g)$. In this case one lineage was sampled from each species but the species tree and the gene tree differ topologically. (Right) A metric gene tree with the same topology as the gene tree in the left.

A challenge for inference methods arises from the fact that many data sets involving different species exhibit *gene tree incongruence*: some gene trees relate taxa differently. A possible reason for this is *incomplete lineage sorting* (ILS) [CKC07; Pol+06; Syr+05]. ILS occurs when a lineage divides, within a population, into two new lineages in a way that the resulting lineage ancestry does not depict the same story as the species tree. Figure 3 depicts ILS, where the species tree in the left has a gene tree inside showing the ancestry of the gene sampled from different species. This gene tree has the same topology as the tree in the right of Figure 3 where it is drawn without self-intersections. For instance, it is estimated that the gene trees depicting the same story as the species tree of human, chimp and gorilla $((h, c), g)$ are roughly 78%; around 11% depict the same story as the tree $((h, g), c)$ and around 11% depict $((c, g), h)$ [Ebe+07]. ILS is modeled by the multispecies coalescent model (MSC) [PN88], which we describe in Section 2. See [Wak08] for a survey of references on the MSC.

Software has been developed to simulate the multispecies coalescent process on a species tree. Multispecies coalescent simulators will generate a set of metric gene trees when a metric species tree with population size functions are given as input (usually simulators will only allow constant population functions). These simulators aid researchers testing their inference methods in the presence of ILS; thus it is important for these simulators to be accurate. While these simulators are usually tested, no thorough testing procedure guarantees their accuracy. Here we introduce a method of testing coalescent simulators using the exact probability density function of the distance between two lineages, each sampled from two different species, to see if it matches the pairwise distance between these lineages obtained by the simulated metric gene trees. We then apply the method to four coalescent simulators.

2. Background

2.1. **The Coalescent Model: Single population.** Coalescent theory models the formation of gene trees within populations of species. The coalescent model for a single population traces (backwards in time) the ancestries of a finite set of individual copies of a gene as the lineages *coalesce* to form ancestral lineages. Figure 4 depicts the formation of a gene tree in a single population where 3 lineages were sampled. On the left of this figure each horizontal line of dots represents a generation, and each dot represents an individual. We see how each lineage traced backwards in time from an individual to an individual in a previous generation until the first coalescent event occurs at the 5th generation. The second coalescent event occurs at the 8th generation. On the right of Figure 4 we see the gene tree that was obtained from this process. This depiction of the coalescent is actually a discretization of the coalescent model known as the Wright-Fisher model, with the coalescent model using continuous notions of both time and population sizes.

The coalescent process of lineages is independent of which lineages are present at any specific time — this is known as the *exchangeability property* of the coalescent model. In the coalescent model, the instantaneous rate of coalescence for any two lineages present in a population b with population size N is given by $1/N$. Let $p(t_k)$ be the probability that no two lineages out of k lineages present have coalesced by time t_k in population b. There are $\binom{k}{2}$ possible pairs of lineages and the coalescence of any pair is independent, so the model tells us [AR05]:

$$p'(t_k) = -\binom{k}{2}\frac{p(t_k)}{N_b(t_k)} = -\frac{k(k-1)p(t_k)}{2N_b(t_k)},$$

where $N_b(t)$ is the size of population b at time t.

Thus

$$p(t_k) = \exp\left(-\int_0^{t_k} \frac{k(k-1)}{2N_b(t)}dt\right). \tag{2}$$

Therefore, given k gene lineages in a population b, the probability density of the time t_k until a pair of genes coalesce is given by

$$f(t_k) = \frac{d}{dt_k}(1-p(t_k)) = \frac{k(k-1)}{2N_b(t_k)}\exp\left(-\int_0^{t_k}\frac{k(k-1)}{2N_b(t)}dt\right) = \binom{k}{2}\frac{1}{N_b(t_k)}\exp\left(-\binom{k}{2}\int_0^{t_k}\frac{1}{N_b(t)}dt\right). \tag{3}$$

The instantaneous rate at which coalescent events occur is given by $(\binom{k}{2}/N_b(t_k))^{-1}$. Note that whenever $N_b(t_k) = N$ is constant, $f(t_k)$ is the probability density of an exponential distribution with rate $N/\binom{k}{2}$. Note also that by the exchangeability property of the coalescent model, this derivation can be interpreted as a sum of independent Poisson processes.

When a coalescent event occurs, a new process begins but with one fewer lineage. Every process is only conditioned on the number of lineages present. That is, given a coalescent event between two out of k lineages, the probability density of the time t_{k-1} until the next pair of lineages coalesce is given by equation (2) after replacing all the occurrences of k with $k-1$. See [Wak08] for further detail.

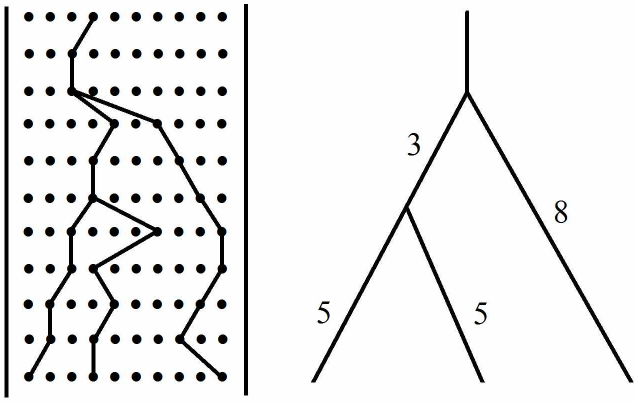

FIGURE 4. (Left) A single population with 3 lineages sampled. Each horizontal line of dots represent a generation, and each dot represent an individual. (Right) The gene tree observed from the coalescent process on the right.

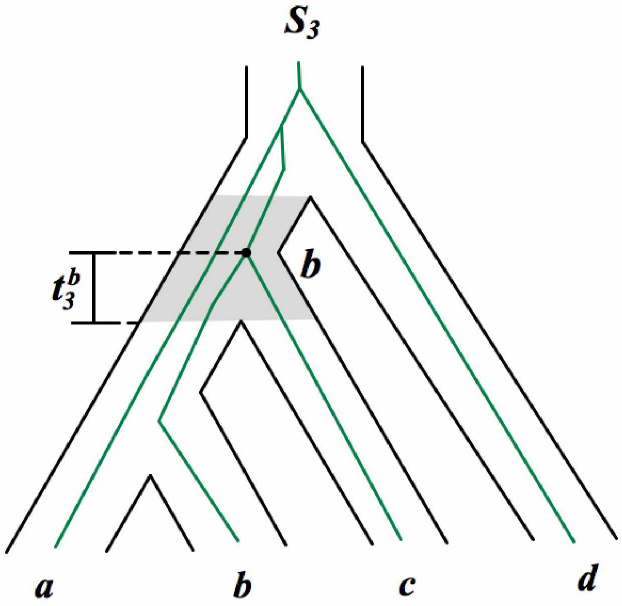

FIGURE 5. A species tree S containing population b (shaded in gray), where 3 lineages enter b and 2 leave it. In this population, t_3^b is the time from entering population b to the coalescent event that reduces the number of lineages from 3 to 2.

2.2. **The Multispecies Coalescent Model (MSC): Multiple populations.** The *multi-species coalescent model (MSC)* is a generalization of the coalescent model, formulated by applying it to multiple populations connected to form a rooted species tree. As mentioned in the introduction, the MSC is commonly used to obtain the probabilities of gene trees in the presence of ILS. Now we briefly explain how to do so. Let b, be a population of length $\tau_b \in (0, \infty)$. Following Chifman and Kubatko [CK15] (who followed [RY03]), let u denote the number of lineages entering b and let v be the number of lineages leaving it. Then there are $u - v$ coalescent events. For example, in Figure 5, there are 3 lineages entering population b (the population shaded in gray) and 2 leaving it, so $u = 3$ and $v = 2$. When there are j lineages present at a certain point

in time, by the exchangeability property of the coalescent model, any of the $\binom{j}{2}$ pairs is equally likely to coalesce. Therefore the density for a coalescent event has to be weighted by $\binom{j}{2}^{-1}$, the probability of a particular pair coalescing. Let t_k^b denote the time from the most recent speciation event that involves b to the coalescent event that reduces the number of lineages in this branch from k to $k-1$. Figure 5 shows an example, where in population b we depict the time t_3^b, which is the time from entering population b to the coalescent event that reduces the number of lineages from 3 to 2. By defining $t_{u+1}^b = 0$, we can write the joint density of coalescent times $t_u^b, t_{u-1}^b, ..., t_{v+1}^b$ within population b as

$$f_b(t_u^b, t_{u-1}^b, ..., t_{v+1}^b) = \prod_{j=v+1}^{u} \left[\frac{1}{2N_b(t)} \exp\left(-\binom{j}{2} \int_{t_{j+1}^b}^{t_j^b} \frac{1}{N_b(t)} dt \right) \right] \exp\left(-\binom{v}{2} \int_{t_{v+1}^b}^{\tau_b} \frac{1}{N_b(t)} dt \right) \tag{4}$$

with $t_j^b \in (0, \infty)$. The probability that there are no coalescent events in b, i.e, $u = v$, is

$$P(\text{no coalescent among } u \text{ lineages in population } b) = \exp\left(-\binom{v}{2} \int_0^{\tau_b} \frac{1}{N_b(t)} dt \right), \tag{5}$$

which follows from equation (2). Equations (4) and (5) describe the coalescent process within a population. When the number of lineages entering and leaving a population on a species tree are specified, the coalescent processes within different populations are conditionally independent. Therefore the probability density for (G, t) on a species tree (S, τ) is given by

$$f((G,t)|(S,\tau)) = \prod_{e=1}^{n-1} f_e(t_{u_e}^e, t_{u_e-1}^e, ..., t_{v_e+1}^e) \tag{6}$$

where the index e is over edges in (S, τ), u_e is the number of lineages entering population e, and v_e is the number of lineages leaving it. A more detailed exposition of this can be found in [Wak08], and [RY03], or in [CK15] for fixed population size.

All these formulas have been derived for *haploid* organisms, this is, organisms with one copy of each gene in their genome. For *diploid* organisms, which have two copies of each gene in their genome, we merely need to replace all the occurrences of $N(t)$ in the formulas with $2N(t)$.

One can see how testing simulators is difficult, given the complicated density and the fact that any metric gene tree that can be embedded in the species tree has positive density.

3. Methods

In this section we present the methods for the two tests developed.

3.1. **Pairwise distance probability density.**

To obtain the pairwise distance probability density between two taxa, we first need the density function of the time to coalescence of two taxa within a single population. This is shown in [ALR18].

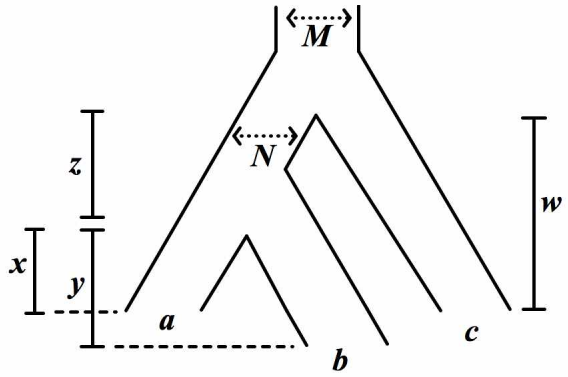

FIGURE 6. The species tree $((a{:}x, b{:}y){:}z\#N, x{:}w)\#M$.

Lemma 1 ([ALR18]). *Suppose two lineages enter a population at $t = 0$ with size function $N(t)$, $0 \leq t < \infty$. Then the time to coalescence has probability density*

$$c(t) = c(N; t) = \ell(t) \exp(-L(t))$$

where $\ell(t) = 1/N(t)$ is the inverse population size and $L(t) = \int_0^t \ell(\tau) d\tau$ its integral.

Observe that when $k = 2$ this lemma follows from equation (3).

On a gene tree, the distance between two lineages A and B (usually denoted by $d(A, B)$) is the sum of number of generations from the present to when lineages A and B enter the same population on the species tree plus twice the time to coalesce. Using Lemma 1 we can compute the pairwise taxon distance distribution. That is, given two lineages A and B, the pairwise distance is the sum of the number of generations where A and B are in different populations (since there cannot be a coalescent event) plus twice the time to coalescence of such lineages (provided they are in the same population). We show that this enables us to find the probability density of $d(A, B)$ with an example before working in a more general setting.

Example 1 Consider a species tree $S = ((a{:}x, b{:}y){:}z\#N, c{:}w)\#M$, as depicted in Figure 6. We do not specify populations on the *pendant edges* (edges incident to the leaves) since we will only sample one lineage from each species and thus there is no coalescent event in such populations. We sample lineage A from species a, B from species b, and C from species c. We first compute the distribution of the pairwise distance of lineages A and C. Observe that these lineages cannot coalesce below the root since they are in different populations until then. The distance $d(A, C)$ is $x + z + w + \epsilon$, where $\epsilon = 2X$ with $X \sim \text{Exp}(1/M)$. Note that $x + z$ is the number of generations from the root to a, w is the number of generations from the root to c, and X has the same probability density function of a 2-lineage coalescent process in a single population with population size M, obtained using Lemma 1. Thus the probability density function for $\ell = d(A, C)$

is

$$f(\ell) = \begin{cases} 0 & \text{if } \ell < x + z + w, \\ \frac{1}{2M} \exp\left(-\frac{\ell - x - z - w}{2M}\right) & \text{if } x + z + w \leq \ell. \end{cases}$$

Analogously, we can compute the probability density function for $\ell = d(B, C)$, as given by the same formula after substituting y for x. The last density, that of $\ell = d(A, B)$, differs from previous ones and is given by a piecewise formula. Each piece where the density is non-zero corresponds to a population where it is possible that A and B can coalesce. In the population immediately above a and b, where the population size is N, $\ell = x + y + 2\epsilon$, where ϵ is drawn from a truncated Exp$(1/N)$. Alternatively, there could be no coalescent event in this edge (which, by equation (5), occurs with probability $\exp(-\frac{z}{N})$) and the lineages will coalesce in the population above the root where the population size is M. In the latter case $\ell = x + y + 2z + 2\epsilon$, where ϵ is drawn from an exponential distribution with parameter $1/M$ down-weighted by $\exp(-\frac{z}{N})$ to reflect the conditioning on no earlier coalescence. Thus the probability density function for ℓ is given by:

$$f(\ell) = \begin{cases} 0 & \text{if } \ell < x + y, \\ \frac{1}{2N} \exp\left(-\frac{\ell - x - y}{2N}\right) & \text{if } x + y \leq \ell < x + y + 2z, \\ \exp(-\frac{z}{N}) \frac{1}{2M} \exp\left(-\frac{\ell - x - y - 2z}{2M}\right) & \text{if } x + y + z + 2z \leq \ell. \end{cases}$$

In this case this density has 3 pieces, one piece, where the density is 0, corresponds to the populations in which these lineages are separated, and the other two pieces correspond to the two different populations these lineages can traceback together.

In the general case, let $(S, \underline{\tau})$ be a metric species tree where each edge has been assigned a population size function, and let a, b be two species in S. Let A be a lineage sampled from a and B be a lineage sampled from b. Let $c(t)$ be the probability density function of the time to coalescence of A and B, as described in Lemma 1. Let k be the *most recent common ancestor* of a and b (that is the node in S where a and b are in the same population for the first time), and let P be the path in S from r, the root of S, to k (since S is a tree, P is uniquely determined [AR05]). Let p be the number of edges in P and let $e_1, e_2, ..., e_p$ be the edges of P, where r is incident to e_1 and k is incident to e_p, and let δ_i be the length of the branch e_i. Figure 7 shows an example of a species tree which for species a and b has $p = 4$. Let $N_i(t)$ be a population size function associated with e_i and let $N_r(t)$ be the population size function that occurs above the root. Finally, let g_a and g_b be the number of generations from k to a and b respectively. Then a distance $d(A, B)$ is $g_a + g_b + \epsilon$, where $\epsilon = 2X$ and X has probability density function $c(t)$. Then the probability density for

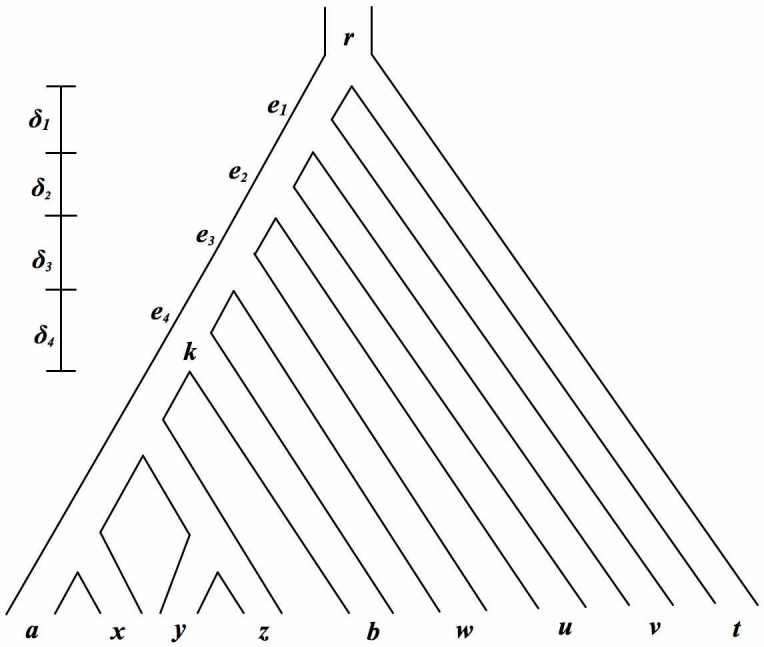

FIGURE 7. A species tree with root r, and where the most recent common ancestor of a and b is labeled by k. The path P is composed of the edges e_1, e_2, e_3, and e_4. Each of these edges has length δ_1, δ_2, δ_3, and δ_4 respectively.

$\ell = d(a,b)$ is given by:

$$f(\ell) = \begin{cases} 0 & \ell \leq g_a + g_b, \\ \frac{1}{2N_p^*(\ell/2)} \exp\left(-\int_{g_a+g_b}^{\ell/2} \frac{1}{N_p^*(t)} dt\right) & g_a + g_b \leq \ell < g_a + g_b + 2\delta_p, \\ \eta_p \frac{1}{2N_{p-1}^*(\ell/2)} \exp\left(-\int_{g_a+g_b+2\delta_p}^{\ell/2} \frac{1}{N_{p-1}^*(t)} dt\right) & g_a + g_b + 2\delta_p \leq \ell < g_a + g_b + 2\delta_p + 2\delta_{p-1}, \\ \eta_{p-1} \frac{1}{2N_{p-2}^*(\ell/2)} \exp\left(-\int_{g_a+g_b+2\delta_p+2\delta_{p-1}}^{\ell/2} \frac{1}{N_{p-2}^*(t)} dt\right) & g_a + g_b + 2\delta_p + 2\delta_{p-1} \leq \ell < g_a + g_b + 2\delta_p + 2\delta_{p-1} + 2\delta_{p-2}, \\ \quad \vdots & \quad \vdots \\ \eta_2 \frac{1}{2N_1^*(\ell/2)} \exp\left(-\int_{g_a+g_b+2\delta_p+\cdots+2\delta_2}^{\ell/2} \frac{1}{N_1^*(t)} dt\right) & g_a + g_b + 2\delta_p + \cdots + 2\delta_2 \leq \ell < g_a + g_b + 2\delta_p + \cdots + 2\delta_1, \\ \eta_1 \frac{1}{2N_r^*(\ell/2)} \exp\left(-\int_{g_a+g_b+2\delta_p+\cdots+2\delta_1}^{\ell/2} \frac{1}{N_r^*(t)} dt\right) & g_a + g_b + 2\delta_p + \cdots + 2\delta_1 \leq \ell, \end{cases}$$

where

$$\eta_j = \eta_p \prod_{i=j}^{p-1} \exp\left(-\int_{g_a+g_b+\sum_{l=i+1}^{p}\delta_l}^{g_a+g_b+\sum_{l=i}^{p}\delta_l} \frac{1}{N_i^*(t)} dt\right)$$

for $j < p$, $\eta_p = \exp\left(-\int_{g_a+g_b}^{g_a+g_b+\delta_p} \frac{1}{N_p^*(t)} dt\right)$, and $N_j^*(\ell)$ has the same range as $N_j(t)$ but the domain is shifted and scaled accordingly to each piece of $f(\ell)$. For example, $N_p^*(\ell/2) = N_p\left(\frac{\ell-(g_a+g_b)}{2}\right)$.

Recall that a probability density function $g(x;\theta)$ belongs to the one parameter exponential family if and only if

$$g(x;\theta) = a(\theta)h(x)\exp\{b(\theta)R(x))\},$$

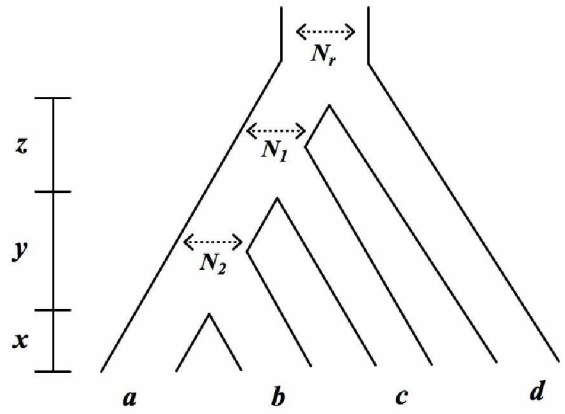

FIGURE 8. A metric species tree $(((a{:}x, b{:}x){:}y, c){:}z, d)$, where the internal branches have population size functions $N_2(t)$, $N_1(t)$, and $N_r(t)$.

where $a(\theta)$, $h(x)$, $b(\theta)$, and $R(x)$ are known functions. Let $L_i(t)$ be the antiderivative of $l_i(t) = 1/2N_i^*(t)$ and let θ be the lower bound of the domain of $N_i^*(\ell)$; that is $\theta = g_a + g_b + \sum_{j=i+1}^{p} 2\delta_j$. Then each piece of $f(\ell)$ is a truncated element of the exponential family weighted by the probability of no coalescence until lineages arrive to certain population (determined by the domain of the $N_i^*(t)$), where

$$a(\theta) = \exp\{L_i(\theta)\},$$

$$h(\ell) = 1/2N_i^*(\ell),$$

$$b(\theta) = 1,$$

$$R(\ell) = -L_i(\ell).$$

In particular note that when $N_i^*(t)$ is constant, the piece of $f(\ell)$ with the corresponding domain is a scaled, truncated exponential distribution (as shown in Example 1). When $N_i^*(\ell) = (c\ell)^{-p}$ for some $c, p > 0$, the corresponding piece of $f(\ell)$ is a scaled, truncated Weibull distribution. For general $N_i^*(\ell)$, the corresponding piece of $f(\ell)$ does not fall into any well-known family.

We developed a function in R [R C13] called `pairwisedist.r` that computes $f(\ell)$. This function takes as input a species tree topology S, edge lengths $\lambda : E(S) \to \mathbb{R}_{\geq 0}$, a population function for each edge and the root (already in the form $N_i^*(\ell)$), and two species of S. The output of this function is the probability density of the distance of two lineages sampled for the species. For example, let S be the species tree shown on Figure 8, where $x, y, z = 1000$, $N_r^* = \frac{t}{2} + 1000$, $N_2^* = 2t + 1000$, and $N_3^* = 2000$. Then `pairwisedist.r` produces the values depicted in Figure 9. The function is 0 between 0 and 2000 because the populations where these lineages are apart have 1000 generations each. The discontinuities of the density function correspond to the changes in population sizes.

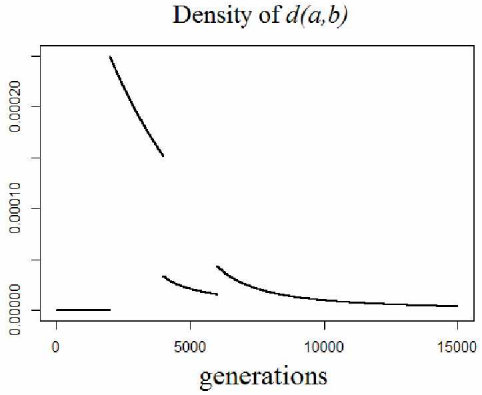

FIGURE 9. The plot of the probability density function of $d(a, b)$ in the tree shown in Figure 8.

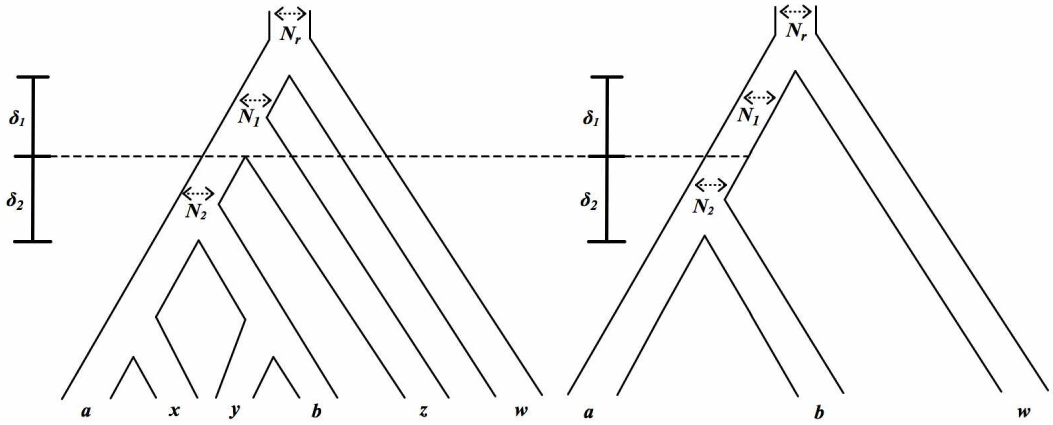

FIGURE 10. (Left) A species tree S with population sizes N_1, N_2 and N_r. (Right) The triplet induced by a, b and w of the species tree on the left.

3.2. **Gene tree topology counts.** Let (S, τ) be a metric species tree and let a, b, and c be species of S. The *triplet induced by a, b, and c*, denoted by S_{abc}, is the tree obtained after removing all populations ("tubes") where neither of a , b, nor c can traceback. In a species tree with n species there are $\binom{n}{3}$ triplets. In each triplet we refer to the population that is above only two species as the internal branch. Figure 10 shows an example of a species tree S and its induced triplet S_{abw}. In this example the internal branch of S_{abw} is above species a and b, its branch length is $\delta_1 + \delta_2$, and its population size function is given by

$$N(t) = \begin{cases} N_2 & \text{if } t < \delta_2, \\ N_1 & \text{if } \delta_2 \leq t < \delta_1. \end{cases}$$

There are three possible gene tree topologies involving lineages A, B and W. These are

$$((A, B), W), \quad ((A, W), B), \quad ((B, W), A).$$

Using the techniques in [PN88], one can show that

$$1-\frac{2}{3}\exp\left(-\left(\frac{\delta_2}{N_2}+\frac{\delta_1}{N_1}\right)\right),\quad \frac{1}{3}\exp\left(-\left(\frac{\delta_2}{N_2}+\frac{\delta_1}{N_1}\right)\right),\text{ and }\frac{1}{3}\exp\left(-\left(\frac{\delta_2}{N_2}+\frac{\delta_1}{N_1}\right)\right) \tag{7}$$

are the probabilities of observing these gene tree topologies respectively. Observe that the probability of the gene trees that do not match the species tree are equal. This is in fact true for any induced triplet of an arbitrary species tree. The internal branch has length $\frac{\delta_2}{N_2}+\frac{\delta_1}{N_1}$ in *coalescent units*, which are the units obtained by scaling time in number of generations inversely by population size. The probability of each gene tree topology only depends on the length in coalescent units of the internal branch. In an induced triplet, the population size and edge length of the internal branch are not identifiable individually but the coalescent units are. This is relevant for us since we can provide an estimate in the coalescent units of the length on the species tree of an internal branch for any induced triplet, as we do in the results section.

4. Results

4.1. **Pairwise distance test.** We used `pairwisedist.r` to test four well-known multispecies coalescent simulators: Mesquite [MM18], SimPhy [MDOMP16], Hybrid-Lambda [Zhu+15], and Phybase [LY10]. One would expect that for a sufficiently large sample size (100,000 gene trees) the histogram of the distance of two fixed lineages will approximate the probability density of the distance of such lineages quite closely. For all simulators we tested four metric species trees

$S_1 = ((a{:}1000, b{:}1000){:}1000\#2000, c{:}2000)\#1000,$

$S_2 = (((a{:}1000, b{:}1000){:}1000\#1000, c{:}2000){:}1000\#1000, d{:}3000)\#1000,$

$S_3 = (((a{:}1000, b{:}1000){:}1000\#2000, c{:}2000){:}1000\#3000, d{:}3000)\#1000,$ and

$S_4 = ((((((a{:}1000, b{:}1000){:}1000\#1000, c{:}2000){:}1000\#3000, d{:}3000){:}1000\#2000, e{:}4000){:}1000\#1000), f{:}5000)\#2000,$

as depicted in Figure 11.

Note that in each of S_1, S_2, S_3, and S_4 the number of generations from the root to each of the species (leaves) is the same. A tree that satisfies this property is known as an *ultrametric tree*. These simulations can be easily extended to non-ultrametric trees but we leave this for further work. All the trees considered here are elements of the family of the *caterpillar trees* [AR05]. The simulators to be tested here only admit constant population size in each edge, and the caterpillar trees allow us to have more changes in population size for pairs of species than any other tree with the same number of taxa. For example, the tree $((a,b),(c,d))$ admits at most two changes in population for any pair of species whereas $(((a,b),c),d)$ admits three changes of population size for species a and b.

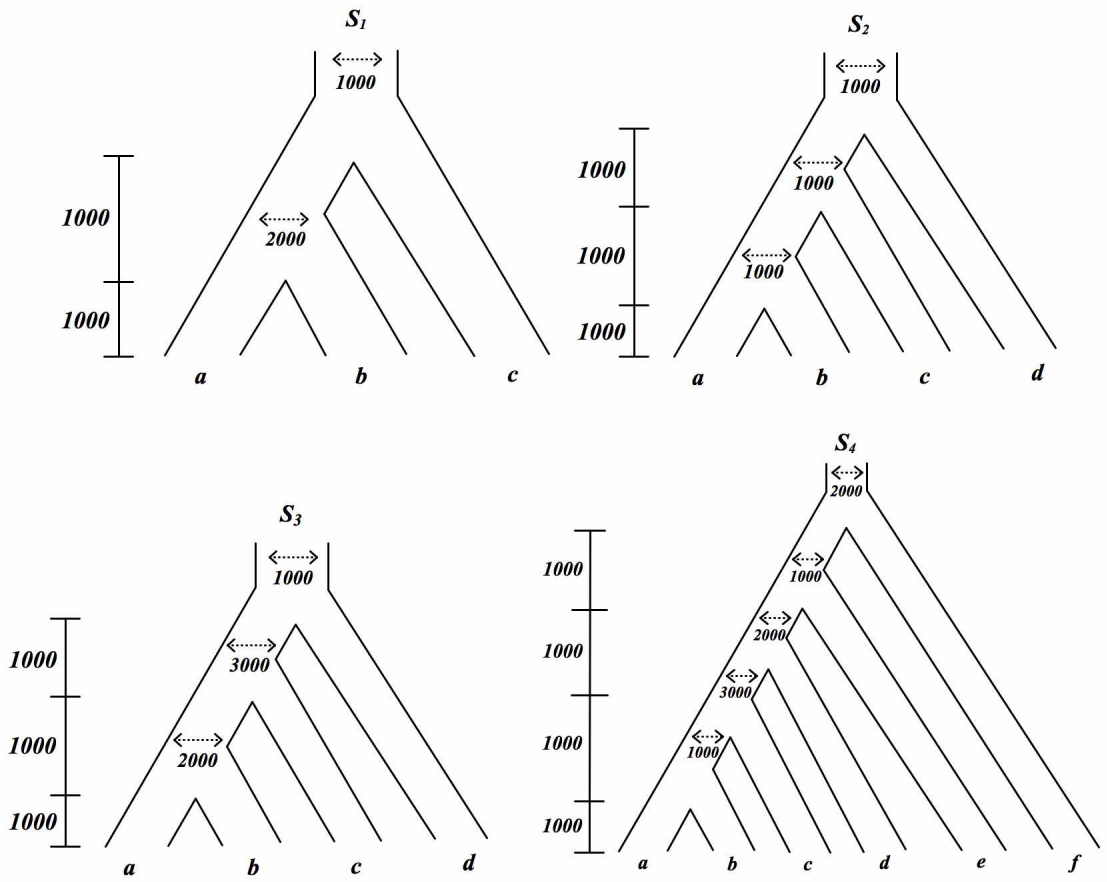

FIGURE 11. The species trees S_1, S_2, S_3, and S_4 used to test multispecies coalescent simulators.

4.1.1. *Mesquite.* Quoting the authors of Mesquite[1] [MM18]:

> *Mesquite is modular, extendible software for evolutionary biology, designed to help biologists organize and analyze comparative data about organisms. Its emphasis is on phylogenetic analysis, but some of its modules concern population genetics, while others do non-phylogenetic multivariate analysis.*
>
> *Mesquite's features for handling gene trees within populations and species trees are relevant for population genetics, phylogeography, and study of gene families. They can also be used by analogy for host-parasite or other studies of associated taxa.*

For species trees S_i, $i \in \{1, 2, 3, 4\}$, we simulated with Mesquite (version 3.5) the coalescent process with one lineage sampled from each species. Using the simulated gene trees we created for each tree, histograms obtained from the pairwise distances of the lineages. We also computed the exact pairwise distance probability distribution obtained from `pairwisedist.r` and overlaid it on the histograms. Figures 12, 13, and 14 show results for S_1, S_2, and S_3 respectively. For S_4 we split these histograms in Figures 15 and 16.

[1] https://www.mesquiteproject.org/

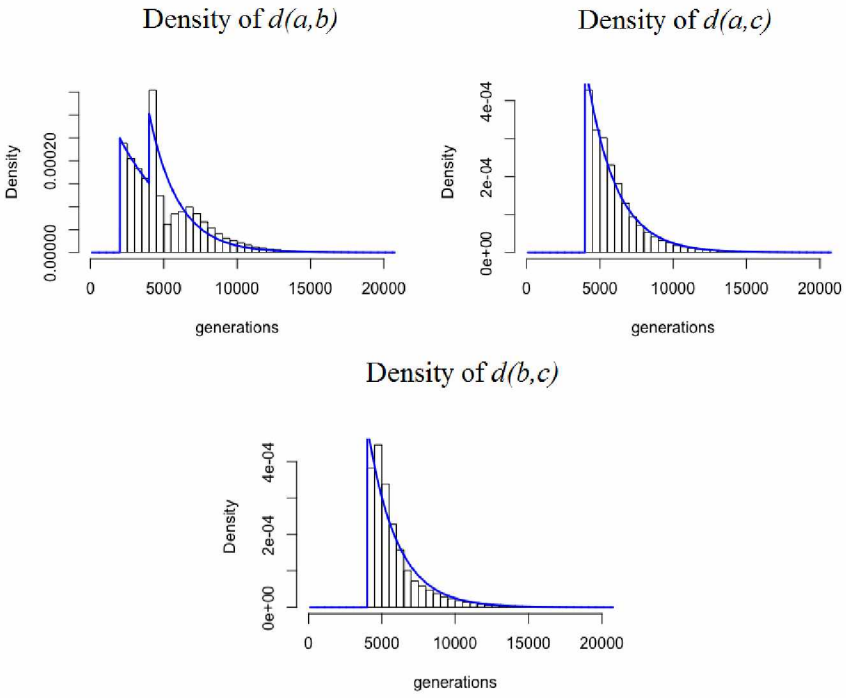

FIGURE 12. The pairwise distance probability distribution of lineages sampled from different species from S_1 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Mesquite.

In Figure 12 we observe that for the histogram of $d(a, b)$ in S_1, Mesquite exhibit problems with its simulations right when the population size changes. At this point, according to Mesquite, lineages are likely to coalesce rapidly followed by a decay steeper than the theoretical one. After this, the probability of coalescence increases again and leads to a exponential decay similar to that predicted by theory. The histogram of $d(a, c)$ fits the theoretical distribution, although we can see that around 6000 generations the histogram does not match the actual distribution closely. By the exchangeability property of the coalescent model the histograms of $d(a, c)$ and $d(b, c)$ should match closely, but they do not. Given the large number of samples we conjecture Mesquite does not behave correctly for this species tree.

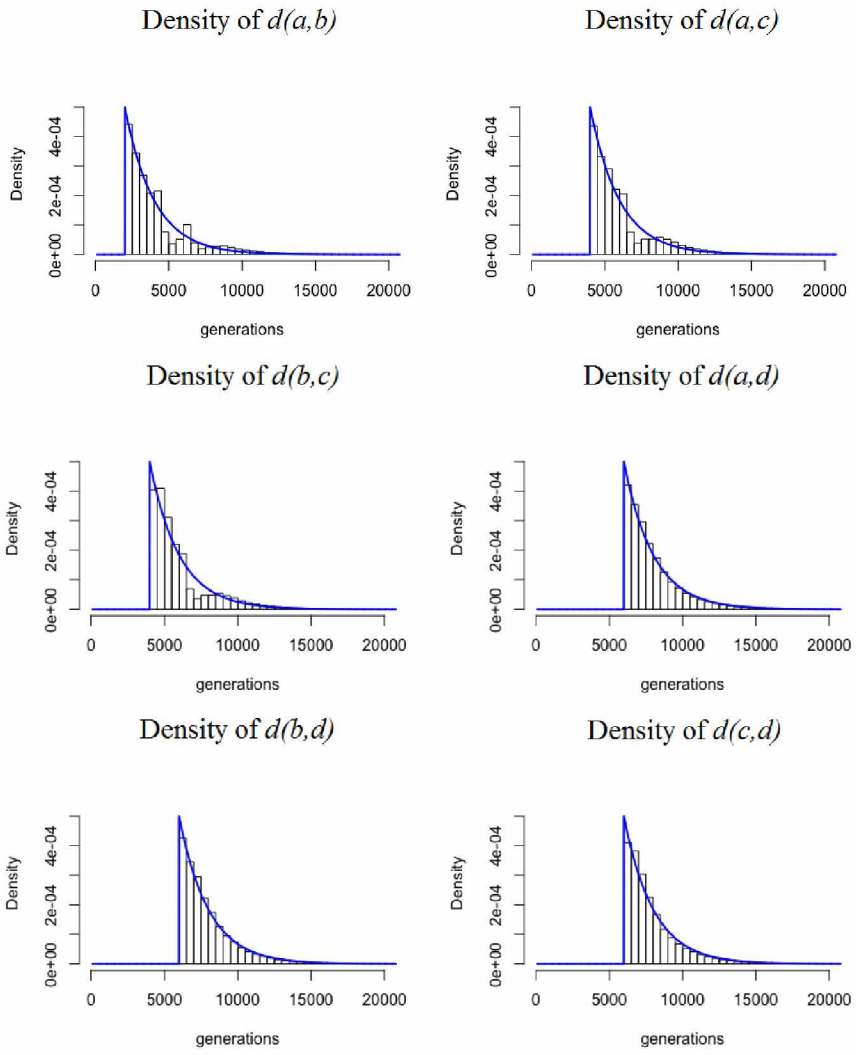

FIGURE 13. The pairwise distance probability distribution of lineages sampled from different species from S_2 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Mesquite.

In Figure 13, which displays the theoretical and simulated density for S_2, a tree with constant population size throughout the tree, we do not see a match in theoretical and simulated empirical distributions in $d(a, b)$, $d(a, c)$, and $d(b, c)$, which should reflect an exponential decay. For the remaining pairwise distances the histograms match the theoretical distribution. Note these distances are the ones with only one population on the species tree relevant to the behavior.

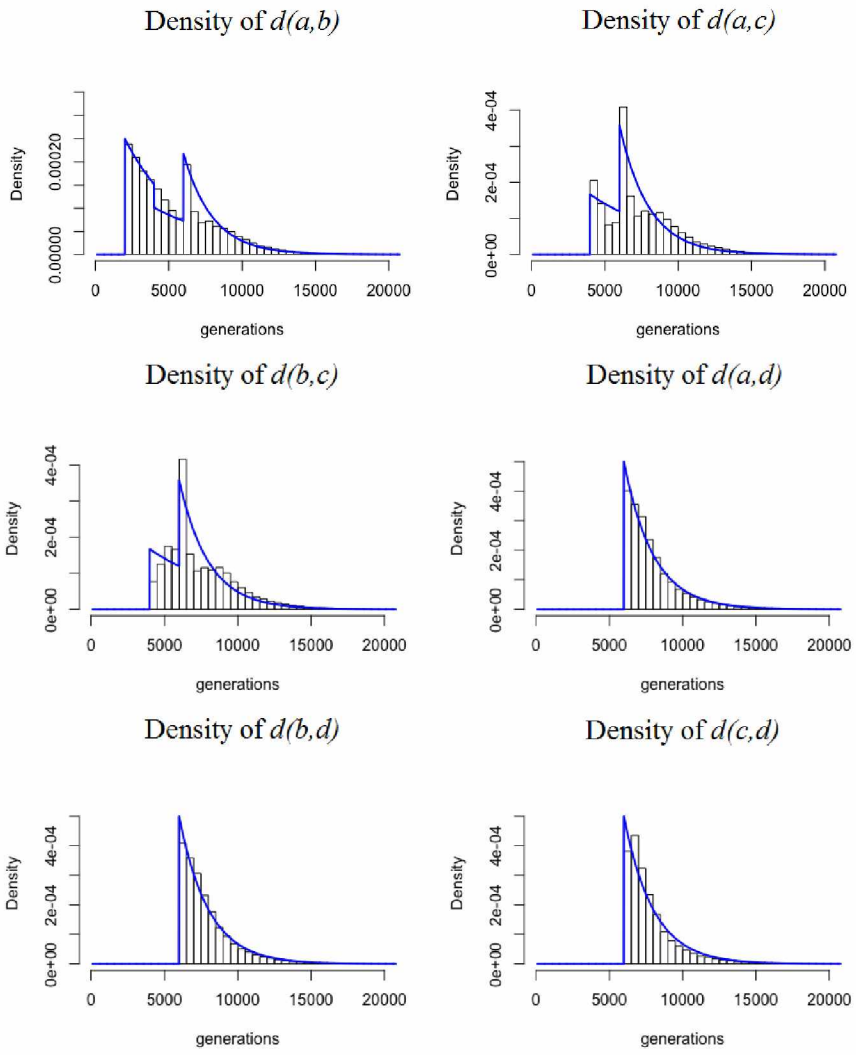

FIGURE 14. The pairwise distance probability distribution of lineages sampled from different species from S_3 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Mesquite.

Turning to Figure 14, the densities for S_3, we do not observe proper behavior in the histograms involving change of population size, and the remaining histograms behave almost as expected. In Figure 15 and 16, which follow, the theoretical and simulated densities of S_4 the results are similar.

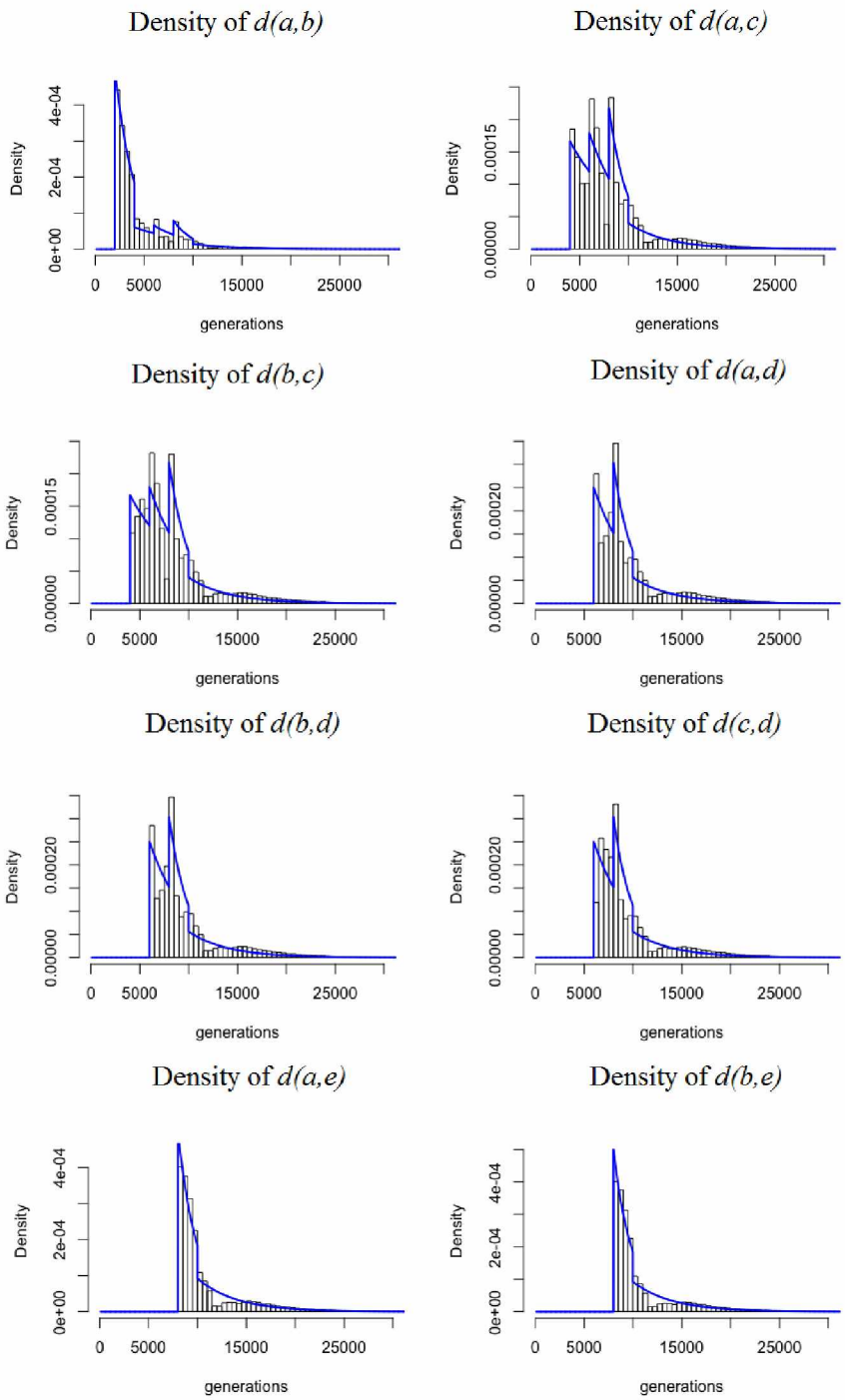

FIGURE 15. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Mesquite.

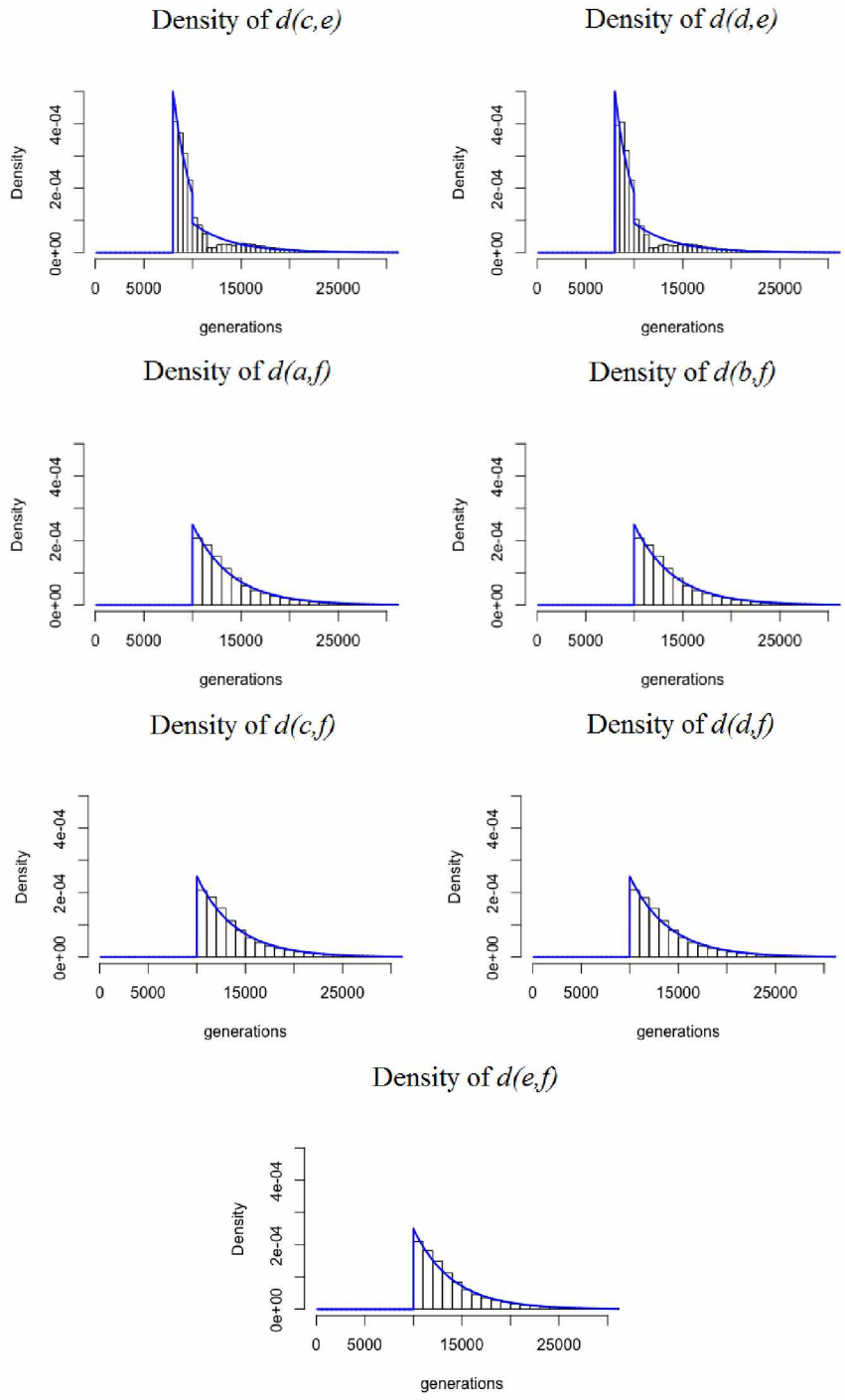

FIGURE 16. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Mesquite.

We conclude that Mesquite simulations do not match the pairwise distributions independently of changes in the population size, or number of species. Any histogram which depends on more than 1 population of the species tree fails to match the theoretical distribution.

4.1.2. *Hybrid-Lambda.* Quoting the authors of Hybrid-Lambda [Zhu+15]:

> *Hybrid-Lambda is a software package that simulates gene genealogies under multiple merger and Kingmans coalescent processes within species networks or species trees. Hybrid-Lambda allows different coalescent processes to be specified for different populations, and allows for time to be converted between generations and coalescent units, by specifying a population size for each population.*

The tests of Hybrid-Lambda (0.6.1-beta (dev)) are analogous to those conduct with Mesquite. Figures 17, 18, and 19 show histograms for S_1, S_2, and S_3 respectively. For S_4 we split these histograms in Figures 20 and 21.

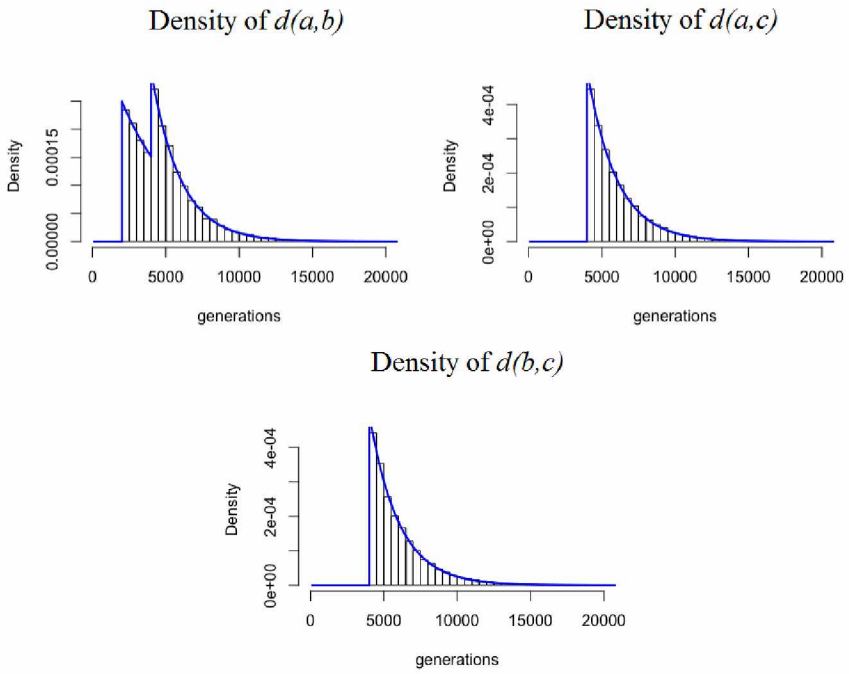

FIGURE 17. The pairwise distance probability distribution of lineages sampled from different species from S_1 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Hybrid Lambda.

We observe that the simulations on S_1, depicted in Figure 17, show the histograms well approximate the probability densities.

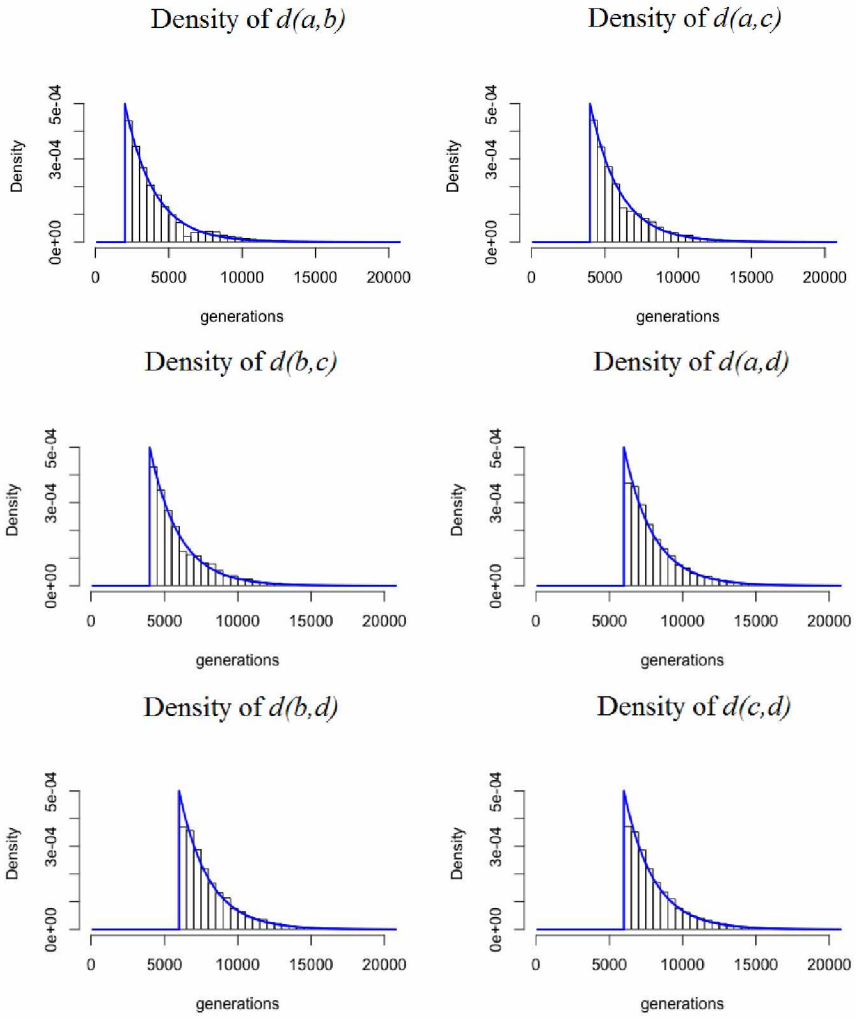

FIGURE 18. The pairwise distance probability distribution of lineages sampled from different species from S_2 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Hybrid-Lambda.

In Figure 18 we observe that the simulated and theoretical densities of S_2 do not match for any distance. The histograms of the densities of $d(a, b)$, $d(a, c)$, and $d(b, c)$ are similar to each other (as they should be by exchangeability) but differ with the theoretical density some time after the lineages enter the same population. The histograms of the densities of $d(a, d)$ and $d(b, d)$ are also similar to each other, as they should be by exchangeability, but show a similar problem as soon as the lineages enter to the same population. Note that the sampled density of $d(c, d)$ shows the same problem as the densities of $d(a, d)$ and $d(b, d)$.

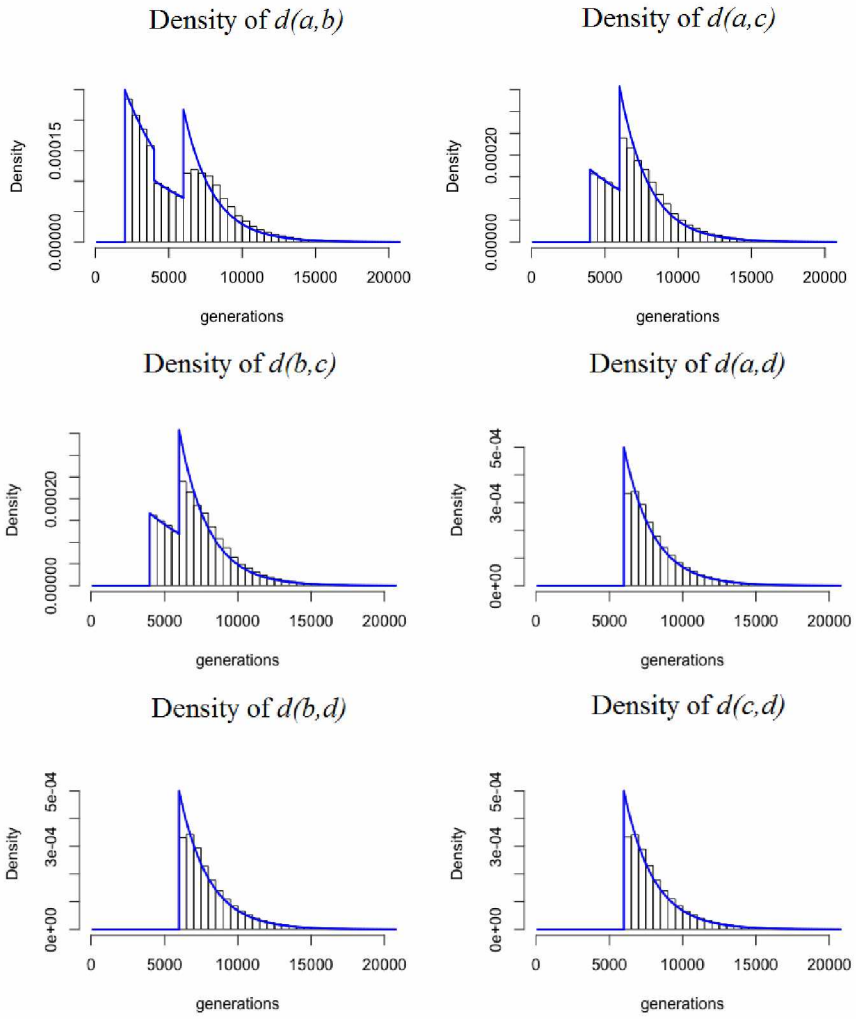

FIGURE 19. The pairwise distance probability distribution of lineages sampled from different species from S_3 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Hybrid-Lambda.

For S_3, we observe in Figure 19 a mismatch between all histograms and the theoretical densities. All pairs of lineages show a mismatch when the lineages enter the population at the root. For S_4, Figures 20 and 21 show a mismatch in the simulated and theoretical densities for all pairs of taxa. However, for both S_3 and S_4 histograms which should agree by exchangeability do so.

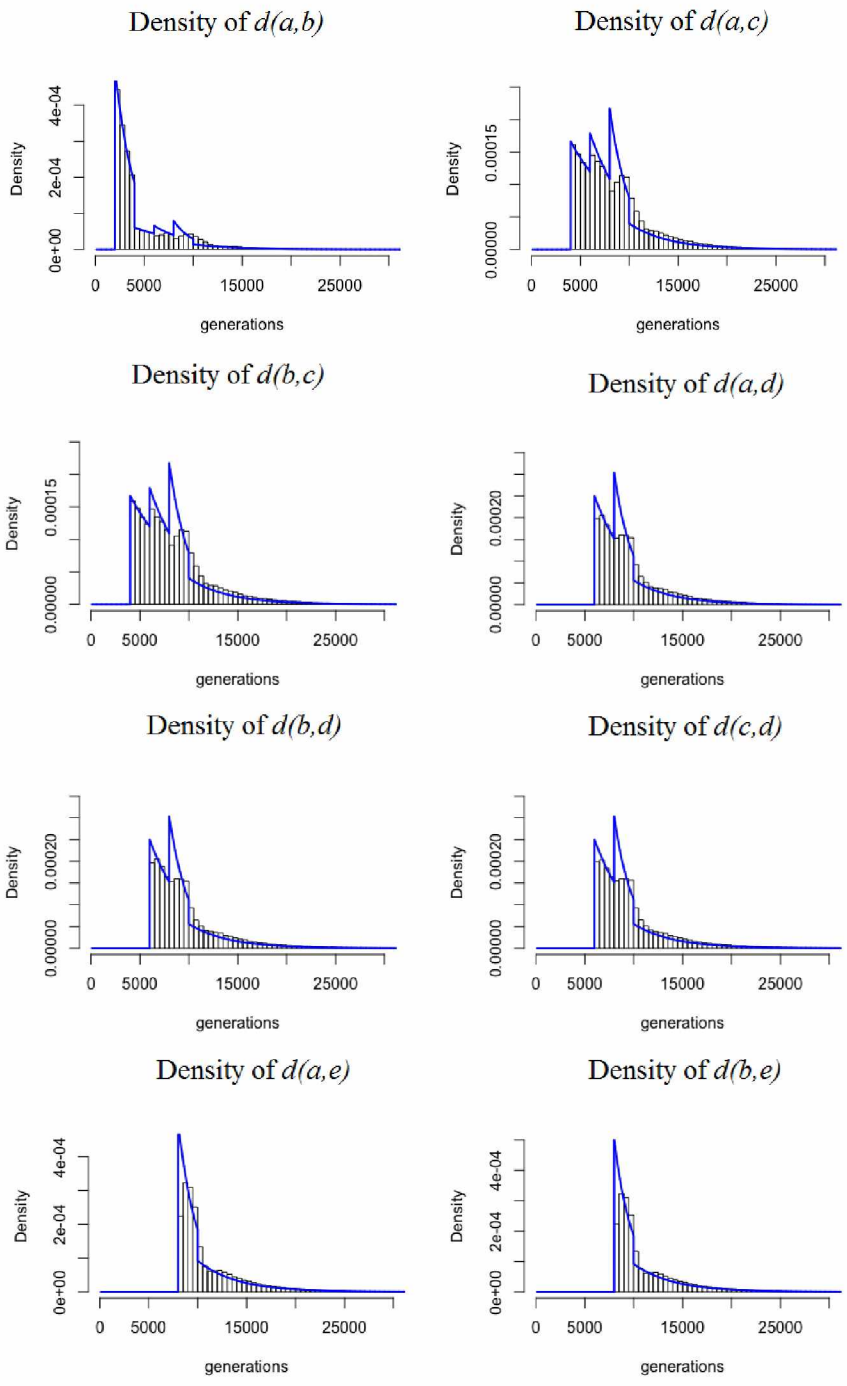

FIGURE 20. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Hybrid-Lambda.

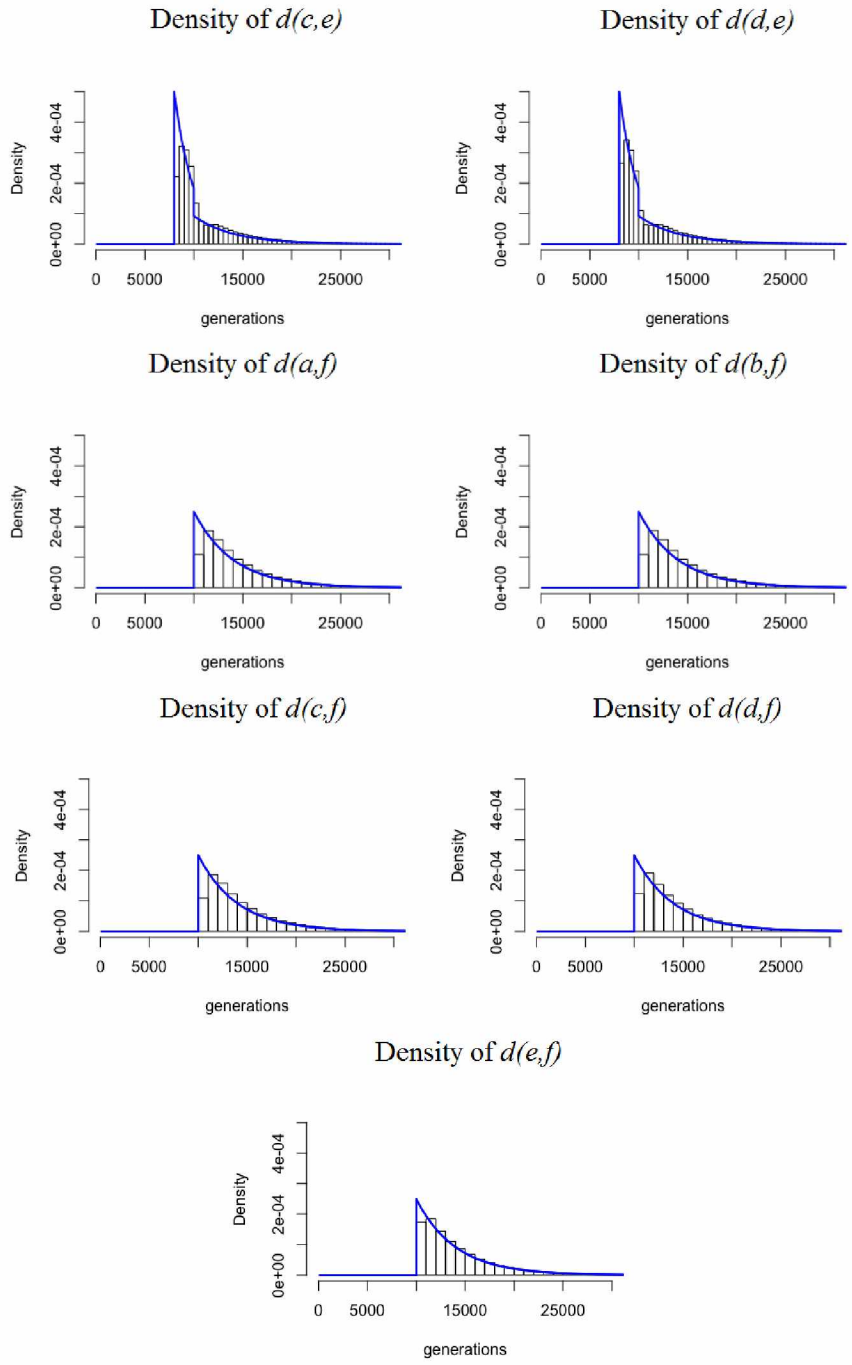

FIGURE 21. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Hybrid-Lambda.

We conclude that Hybrid-Lambda simulations fail to correctly approximate the pairwise distance density with or without changes in the population size for species trees with more than 3 taxa.

4.1.3. *SimPhy.* Quoting the authors of SimPhy [MDOMP16]:

> *We present a fast and flexible software package SimPhy for the simulation of multiple gene families evolving under incomplete lineage sorting, gene duplication and loss, horizontal gene transfer all three potentially leading to species tree/gene tree discordance and gene conversion. SimPhy implements a hierarchical phylogenetic model in which the evolution of species, locus, and gene trees is governed by global and local parameters (e.g., genome-wide, species-specific, locus-specific), that can be fixed or be sampled from a priori statistical distributions. SimPhy also incorporates comprehensive models of substitution rate variation among lineages (uncorrelated relaxed clocks) and the capability of simulating partitioned nucleotide, codon, and protein multilocus sequence alignments under a plethora of substitution models using the program INDELible.*

The tests of SimPhy (version 1.0.2) are analogous to those with Mesquite and Hybrid-Lambda. Figures 22, 23, and 24 show histograms for S_1, S_2, and S_3 respectively. For S_4 we split these histograms in Figures 25 and 26. We observe that for all four trees the histograms closely match theoretical predictions.

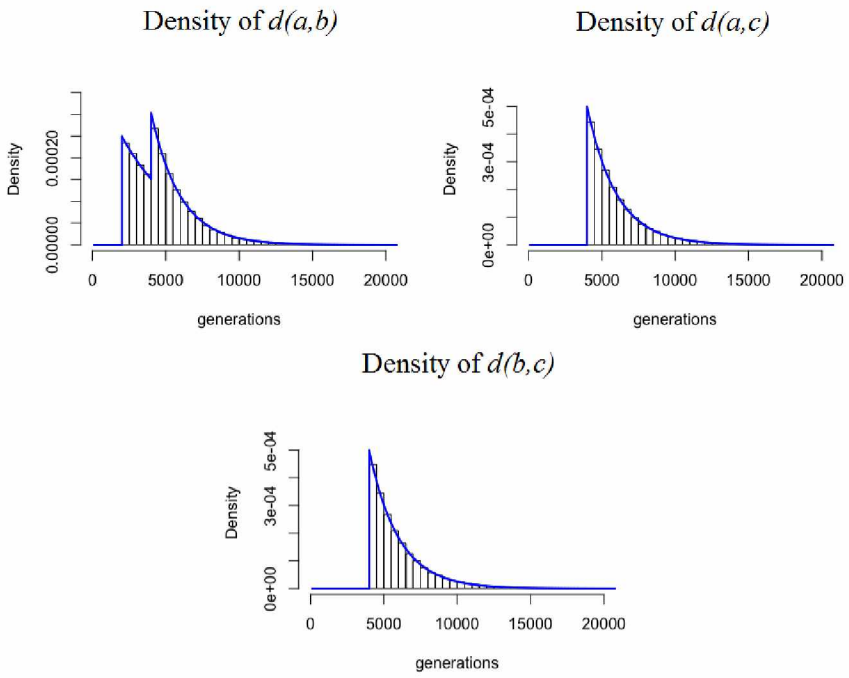

FIGURE 22. The pairwise distance probability distribution of lineages sampled from different species from S_1 together with the histogram obtained from the distances of these lineages from the gene trees simulated by SimPhy.

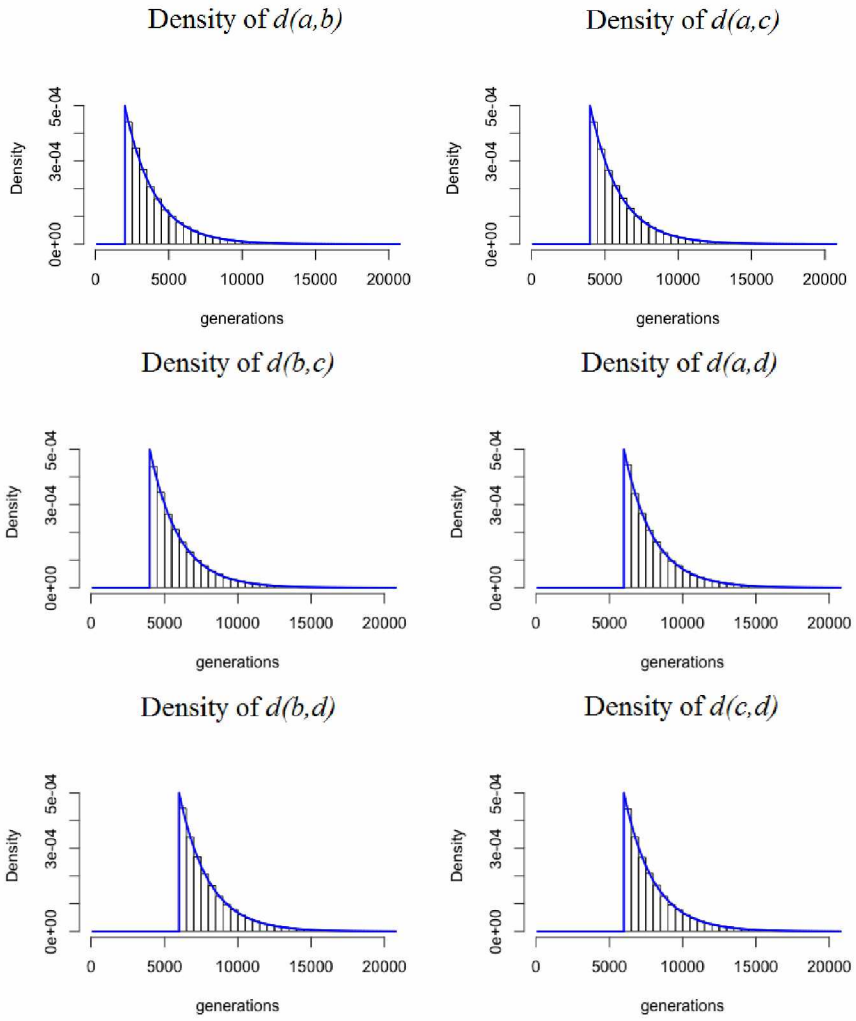

FIGURE 23. The pairwise distance probability distribution of lineages sampled from different species from S_2 together with the histogram obtained from the distances of these lineages from the gene trees simulated by SimPhy.

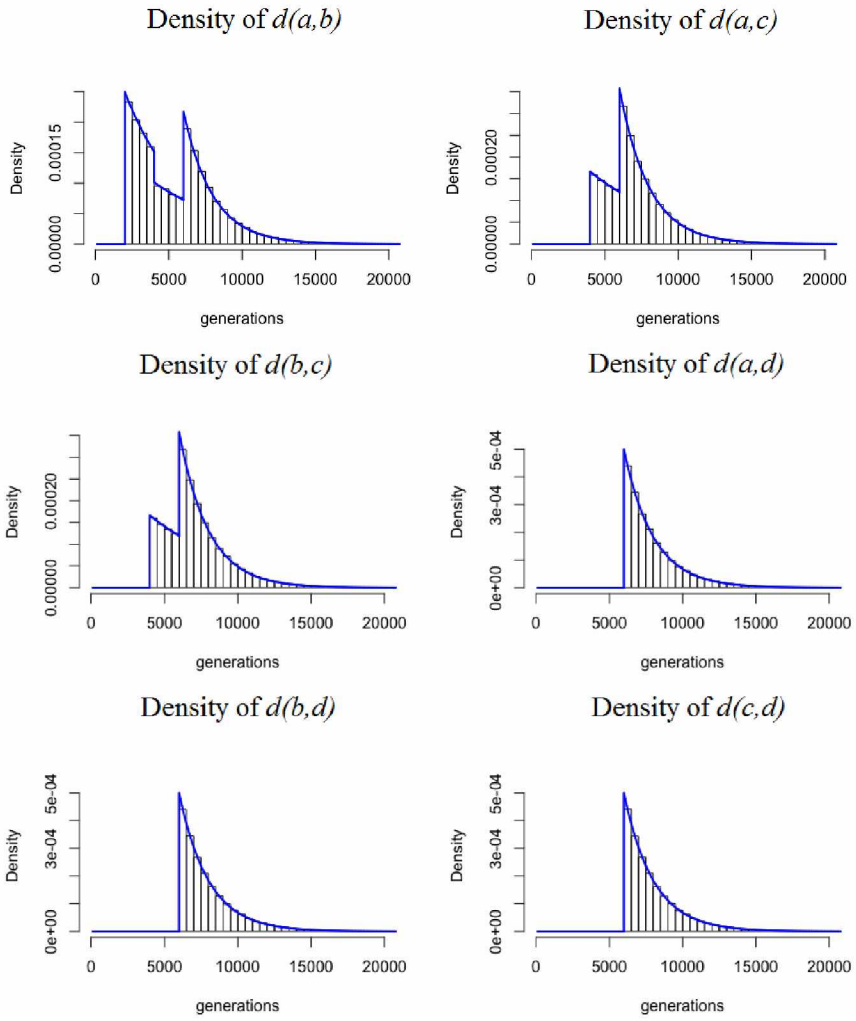

FIGURE 24. The pairwise distance probability distribution of lineages sampled from different species from S_3 together with the histogram obtained from the distances of these lineages from the gene trees simulated by SimPhy.

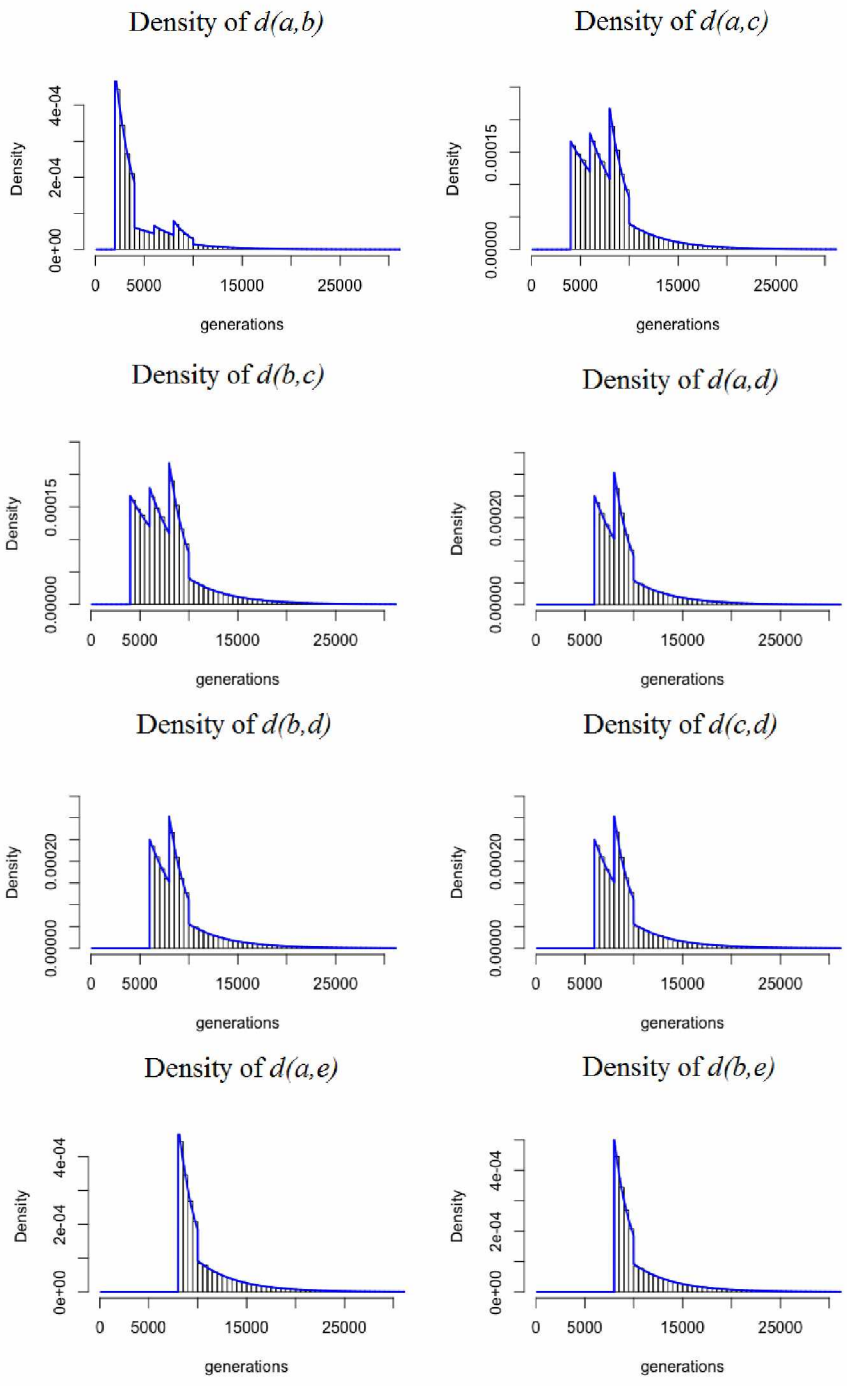

FIGURE 25. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by SimPhy.

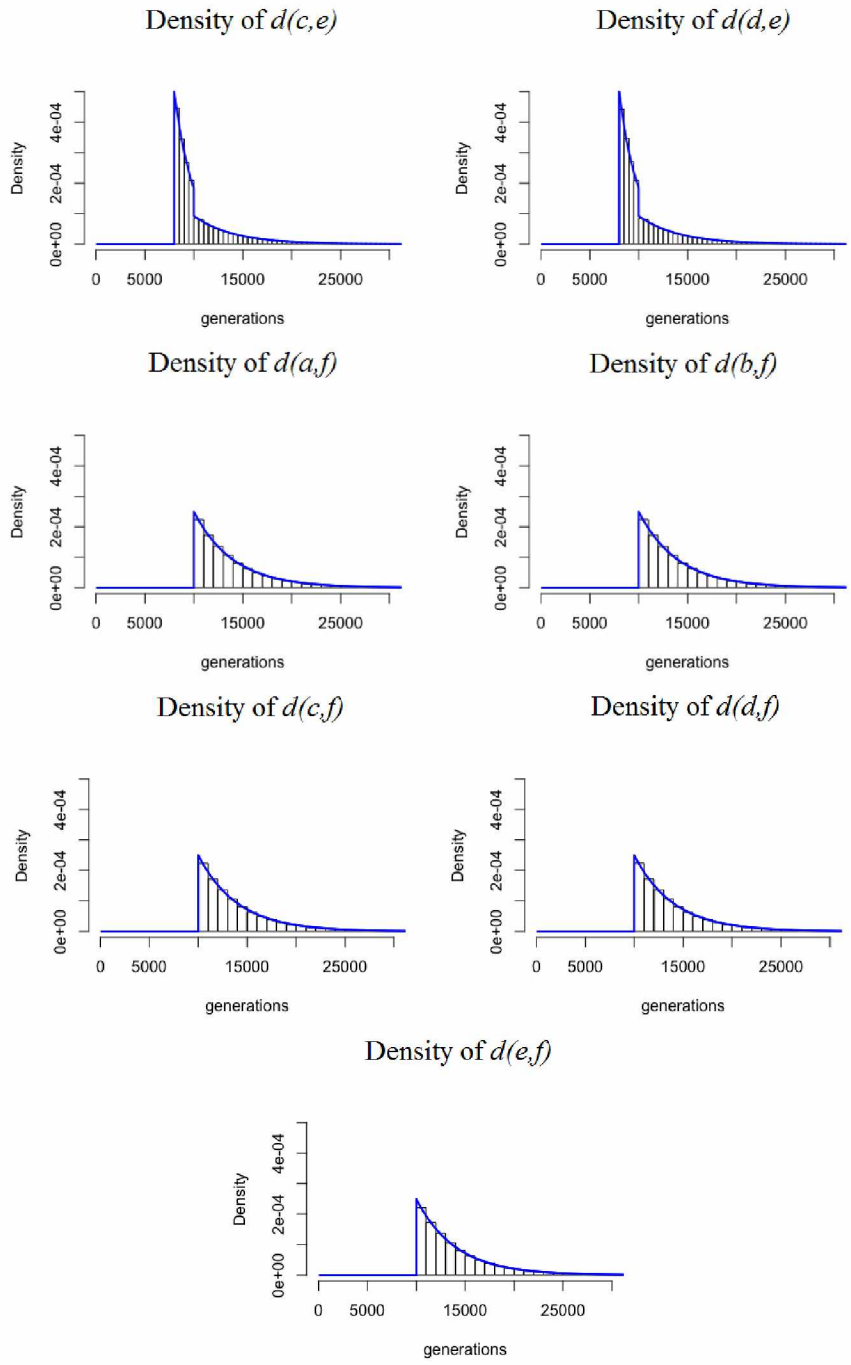

FIGURE 26. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by SimPhy.

4.1.4. *Phybase.* Quoting the author of Phybase [LY10]:

> *Phybase is an R package for phylogenetic analysis using species trees. It provides functions to read, write, manipulate, simulate, estimate, summarize and plot species trees, which contain not only the topology and branch lengths but also population sizes.*

For each species tree S_i, $i \in \{1, 2, 3, 4\}$, we simulated gene trees under the coalescent process with Phybase (version 1.5) with one lineage sampled from each species. If we double the population size in the input to Phybase, the histograms appear as we expect. Clearly, the simulator is not performing correctly for haploid organisms, for otherwise, we should not have to double the population sizes in Phybase's input. Note also, that the simulator is not performing correctly for diploid organisms either, since otherwise, we would have to halve the population sizes in Phybase's input to match theory instead of doubling it. Therefore there is an error in Phybase. Figures 27, 28, and 29 show results for S_1, S_2, and S_3 respectively. For S_4 we split these histograms in Figures 30 and 31. After the adjustment in population sizes, similarly to SimPhy, we observe that for all trees the histograms match the theoretical density.

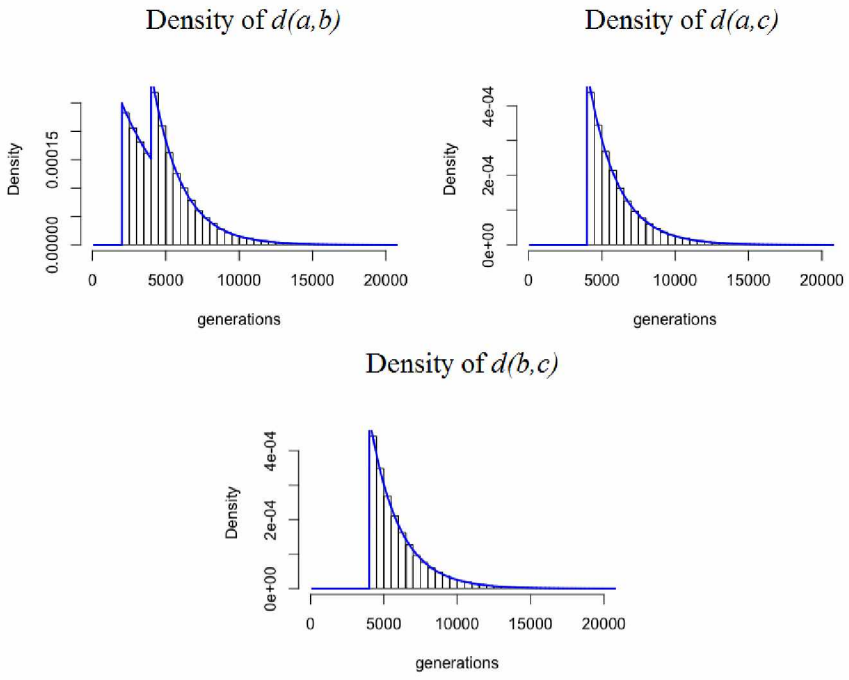

FIGURE 27. The pairwise distance probability distribution of lineages sampled from different species from S_1 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Phybase after adjusting the population sizes.

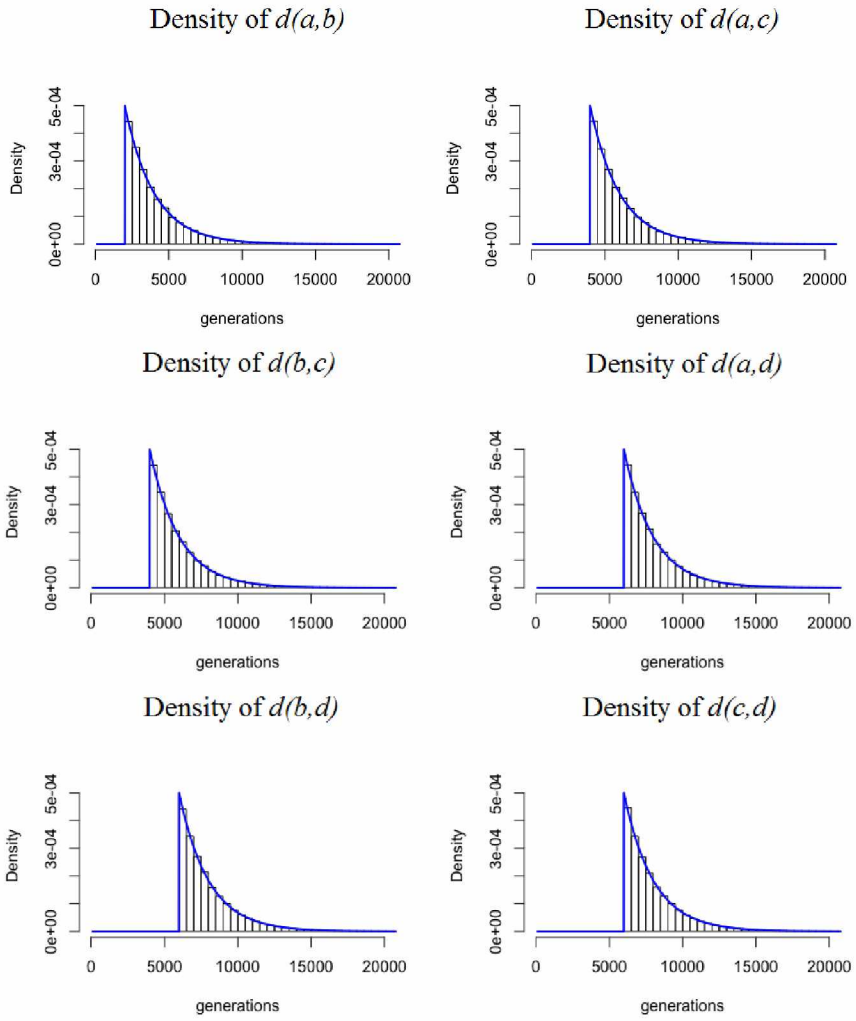

FIGURE 28. The pairwise distance probability distribution of lineages sampled from different species from S_2 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Phybase after adjusting the population sizes.

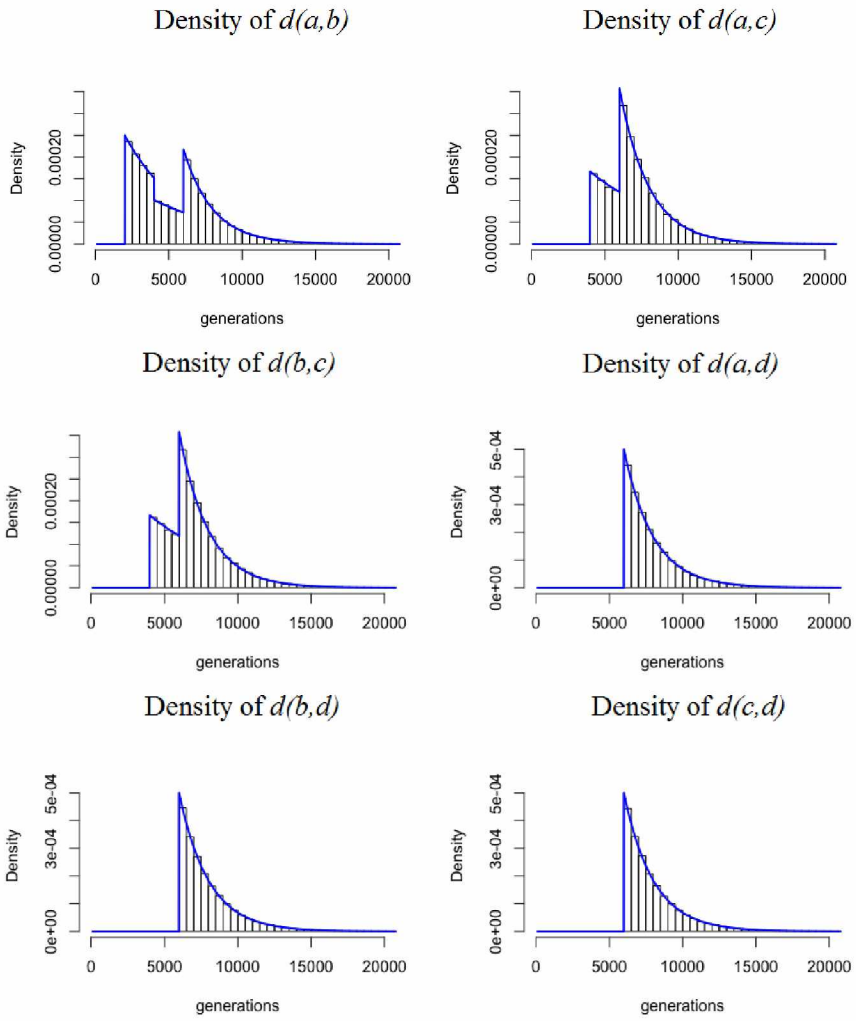

FIGURE 29. The pairwise distance probability distribution of lineages sampled from different species from S_3 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Phybase after adjusting the population sizes.

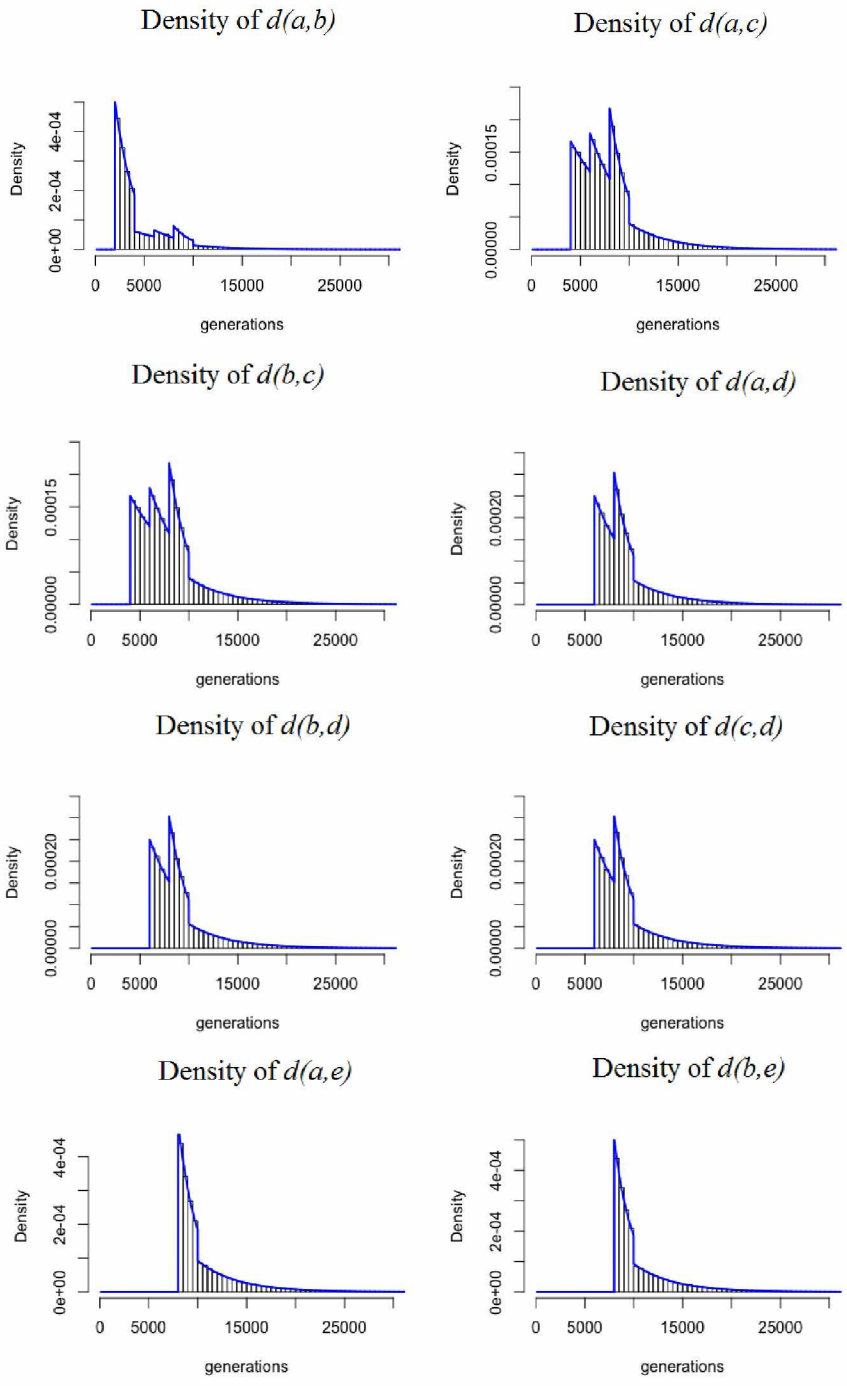

FIGURE 30. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Phybase after adjusting the population sizes.

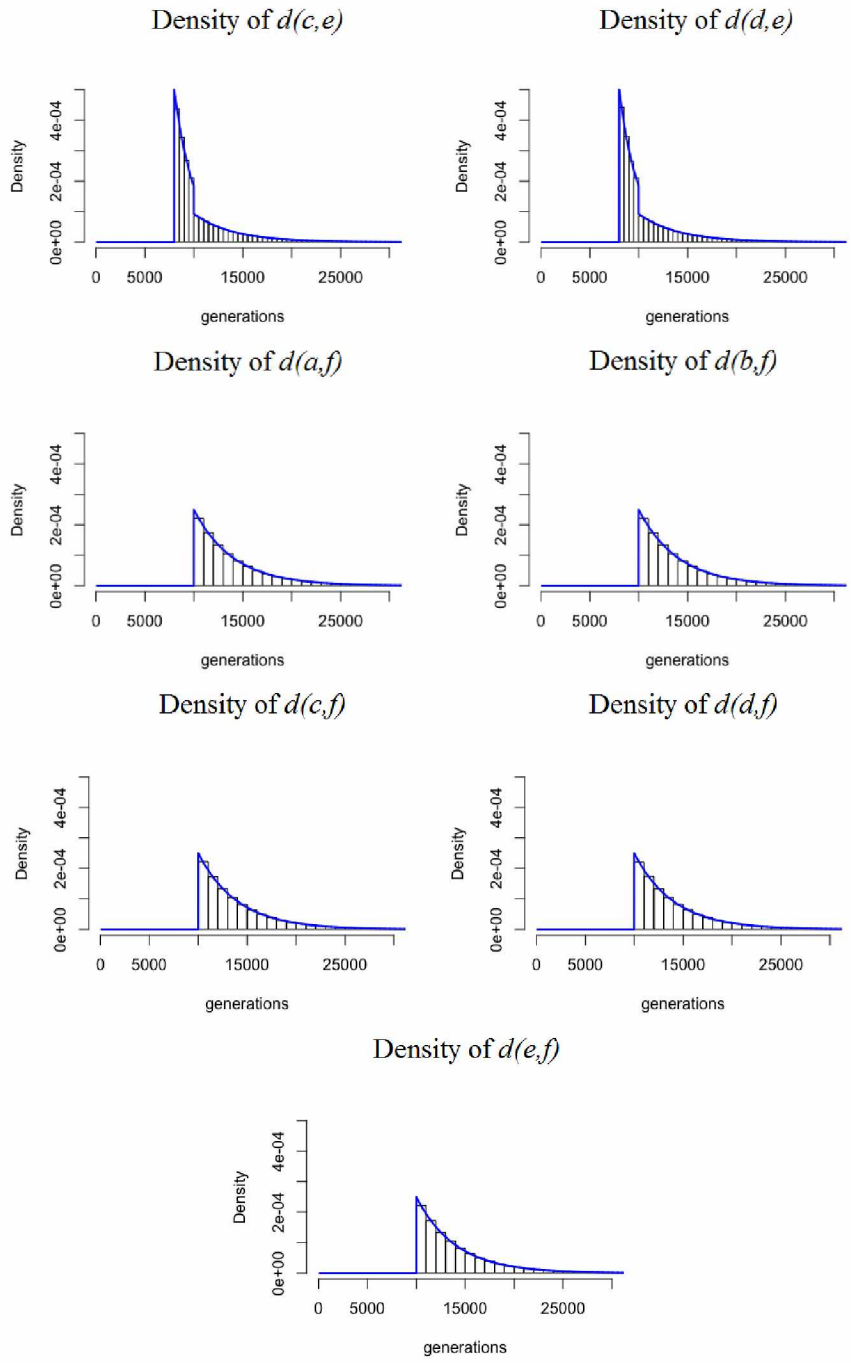

FIGURE 31. The pairwise distance probability distribution of lineages sampled from different species from S_4 together with the histogram obtained from the distances of these lineages from the gene trees simulated by Phybase after adjusting the population sizes.

4.2. **Gene tree topology counts test.** For each of S_1, S_2, and S_3 we examined the set of all possible induced triplets, and for S_4 we choose 5 of the $\binom{6}{3} = 20$ induced triplets. For any given triplet in S_i we computed the probability of observing each of the 3 gene tree topologies as was done in Section 3.2. We multiplied these probabilities by 100,000 (the number of samples) to produce the expected topology count samples. We used the theoretical counts and the simulated counts to perform a χ^2-test with two degrees of freedom. For each gene triplet we used this test to obtain a p-value. These p-values are preliminary results, since this procedure has to be repeated multiple times to obtain an "accurate" conclusion, but we left this for future work. We also computed the internal branch length δ, in coalescent units, for each induced triplet and compared with an estimate $\hat{\delta}$ obtained using the following equation

$$\hat{\delta} = -\log\left(\frac{3}{2} \cdot \frac{T}{100,000}\right) \tag{8}$$

where T is the number of samples whose gene tree topologies do not match the induced triplet topology. Equation (8) is obtained by averaging the last two equations on (7) and solving for the internal branch length in coalescent units. As done in the previous section we doubled the populations sizes in the input to Phybase to obtain the expected behavior in the topology counts.

Tables 1, 2, 3, and 4 show this for the trees S_1, S_2, S_3, and S_4 respectively.

TABLE 1. Topology counts and internal branch estimation for S_1.

Source	$((A,B),C)$	$((A,C),B)$	$((B,C),A)$	p-value	Internal branch
Theoretical	59565	20218	20218	-	0.5
Mesquite	59629	20031	20340	0.281	0.501
Hybrid-Lambda	59511	20131	20357	0.501	0.498
Phybase (adjusted pop. size)	59597	19999	20404	0.128	0.5008
SimPhy	59490	20289	20221	0.841	0.498

In Table 1, where the results for S_1 are shown, we see that all simulators give pretty good estimates of the topology counts and internal branch length in coalescent units. Also, we see no extreme p-values.

TABLE 2. Topology counts and internal branch estimation for the triplets of S_2

Source	$((A,B),C)$	$((A,C),B)$	$((B,C),A)$	p-value	Internal branch
Theoretical	75474	12262	12262	-	1
Mesquite	76009	11964	12027	0.0004	1.022
Hybrid-Lambda	75385	12330	12284	0.770	0.996
Phybase (adjusted pop. size)	75224	12418	12358	0.168	0.989
SimPhy	75497	12266	12237	0.970	1.0009
Source	$((A,B),D)$	$((A,D),B)$	$((B,D),A)$	p-value	Internal branch
Theoretical	90977	4511	4511	-	2
Mesquite	88090	5915	5995	1.609e-221	1.723
Hybrid-Lambda	90834	4525	4640	0.138	1.984
Phybase (adjusted pop. size)	90968	4521	4511	0.988	1.998
SimPhy	90915	4484	4601	0.367	1.993
Source	$((A,C),D)$	$((A,D),C)$	$((C,D),A)$	p-value	Internal branch
Theoretical	75474	12262	12262	-	1
Mesquite	75815	12083	12102	0.044	1.013
Hybrid-Lambda	75574	12271	12154	0.579	1.004
Phybase (adjusted pop. size)	75445	12353	12202	0.612	0.998
SimPhy	75305	12362	12333	0.448	0.993
Source	$((B,C),D)$	$((B,D),C)$	$((C,D),B)$	p-value	Internal branch
Theoretical	75474	12262	12262	-	1
Mesquite	78123	10981	10896	5.09e-83	1.114
Hybrid-Lambda	75579	12163	12257	0.622	1.0042
Phybase (adjusted pop. size)	75323	12215	12462	0.153	0.993
SimPhy	75256	12444	12300	0.178	0.991

In Table 2, where the results for S_2 are shown, we observe that Mesquite gives the worst estimates for all induced triplets. Specially, we see that for the induced triplets $((a,b),d)$ and $((b,c),d)$ the internal branch estimate and p-values are pretty extreme. The other simulators produce good estimates.

TABLE 3. Topology counts and internal branch estimation for the triplets of S_3

Source	$((A,B),C)$	$((A,C),B)$	$((B,C),A)$	p-value	Internal branch
Theoretical	59564	20217	20217	-	0.5
Mesquite	66000	17556	16444	0	0.670
Hybrid-Lambda	59564	20195	20240	0.975	0.499
Phybase (adjusted pop. size)	59395	20256	20349	0.492	0.495
SimPhy	59120	20397	20483	0.554	0.497
Source	$((A,B),D)$	$((A,D),B)$	$((B,D),A)$	p-value	Internal branch
Theoretical	70044	14977	14977	-	$0.83\bar{3}$
Mesquite	74074	12844	13082	4.286e-99	0.944
Hybrid-Lambda	70850	14683	14466	0.207	0.827
Phybase (adjusted pop. size)	71072	14450	14478	0.940	0.834
SimPhy	71079	14465	14456	0.934	0.835
Source	$((A,C),D)$	$((A,D),C)$	$((C,D),A)$	p-value	Internal branch
Theoretical	52231	23884	23884	-	$0.3\bar{3}$
Mesquite	50533	24659	24808	6.16e-26	0.298
Hybrid-Lambda	52230	23951	23818	0.830	0.3331
Phybase (adjusted pop. size)	52419	23606	23975	0.118	0.337
SimPhy	52141	23834	24025	0.579	0.331
Source	$((B,C),D)$	$((B,D),C)$	$((C,D),B)$	p-value	Internal branch
Theoretical	52231	23884	23884	-	$0.3\bar{3}$
Mesquite	51281	24389	24330	1.32e-08	0.313
Hybrid-Lambda	52241	23960	23798	0.758	0.3335
Phybase (adjusted pop. size)	52164	23825	24011	0.635	0.331
SimPhy	52220	23889	23891	0.997	0.3330

In Table 3, where the results for S_3 are shown, we observe that all simulators but Mesquite behave properly. Mesquite behaves poorly for each induced triplet. Similarly, Table 4, which follows, shows the results for S_4.

TABLE 4. Topology counts and internal branch estimation for some randomly chosen triplets of S_4

Source	$((A,B),F)$	$((A,F),B)$	$((B,F),A)$	p-value	Internal branch
Theoretical	96078	1960	1960	-	$2.83\overline{3}$
Mesquite	95535	2275	2190	3.01e-18	2.703
Hybrid-Lambda	96035	1959	2005	0.590	2.822
Phybase (adjusted pop. size)	96055	1997	1948	0.677	2.827
SimPhy	96034	1987	1979	0.749	2.821
Source	$((A,C),F)$	$((A,F),C)$	$((C,F),A)$	p-value	Internal branch
Theoretical	89341	5329	5329	-	$1.83\overline{3}$
Mesquite	86271	6866	6863	8.77e-216	1.58
Hybrid-Lambda	89459	5336	5204	0.212	1.844
Phybase (adjusted pop. size)	89392	5304	5304	0.876	1.838
SimPhy	89403	5304	5293	0.817	1.839
Source	$((B,C),F)$	$((B,F),C)$	$((C,F),B)$	p-value	Internal branch
Theoretical	89341	5329	5329	-	$1.83\overline{3}$
Mesquite	86583	6666	6751	1.88e-174	1.6
Hybrid-Lambda	89298	5330	5371	0.838	1.829
Phybase (adjusted pop. size)	89256	5314	5430	0.361	1.825
SimPhy	89362	5320	5318	0.978	1.835
Source	$((A,D),F)$	$((A,F),D)$	$((D,F),A)$	p-value	Internal branch
Theoretical	85124	7437	7437	-	1.5
Mesquite	80531	9764	9705	0	1.23
Hybrid-Lambda	85152	7540	7307	0.156	1.501
Phybase (adjusted pop. size)	85181	7382	7434	0.800	1.503
SimPhy	84902	7500	7598	0.100	1.485
Source	$((C,D),E)$	$((C,E),D)$	$((D,E),C)$	p-value	Internal branch
Theoretical	59564	20217	20217	-	0.5
Mesquite	61541	19212	19247	6.25e-36	0.55
Hybrid-Lambda	59631	20282	20086	0.567	0.501
Phybase (adjusted pop. size)	59890	20075	20035	0.109	0.508
SimPhy	59701	20256	20043	0.389	0.503

5. Conclusion and discussion

When testing that taxon pairwise distance are accurately simulated, we observe that Hybrid-Lambda and Mesquite exhibit unexpected behavior. We believe Hybrid-Lambda does not have a problem with trees with 3 species, but Mesquite does. We conclude that Mesquite shows bigger discrepancy than Hybrid-Lambda by observing the plots of the tree S_4. We also observe that Phybase, with the correction in population size, and SimPhy exhibit proper behavior.

For the topology count test we see that Mesquite exhibits unexpected behavior when the number of taxa is greater than 4. All the other simulators (and adjusting Phybase's input) seem to match the theoretical

prediction for all cases. This means that one could trust the topology of Hybrid-Lambda even when metric information is incorrect.

More work could be done to assure the validity of the simulations. For example, one could try other types of trees with different number of species.

We could also test different simulators, including those that do not necessarily restrict the population size per edge to be constant. Finally, we could also extend this work for networks, which are an analogous diagram of trees but admit hybridization events. We would compute the pairwise distance probability distribution for lineages in networks and test simulators that allow networks.

6. References

[ALR18] E. Allman, C. Long, and J. Rhodes. "Inferring species trees from genetic sequences using the log-det distance". In: *arxiv.org/abs/1806.04974* (2018).

[AR05] E. Allman and J. Rhodes. "Lecture Notes: The Mathematics of Phylogenetics". In: (2005). URL: https://jarhodesuaf.github.io/PhyloBook.pdf.

[CK15] J. Chifman and L. Kubatko. "Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites". In: *Journal of theoretical biology* 374 (2015), pp. 35–47.

[CKC07] Bryan C. Carstens, L. Lacey Knowles, and Tim Collins. "Estimating Species Phylogeny from Gene-Tree Probabilities Despite Incomplete Lineage Sorting: An Example from Melanoplus Grasshoppers". In: *Systematic Biology* 56.3 (2007), pp. 400–411. ISSN: 1076-836X. URL: http://academic.oup.com/sysbio/article/56/3/400/1655220/Estimating-Species-Phylogeny-from-GeneTree.

[DS05] James H Degnan and Laura a Salter. "Gene tree distributions under the coalescent process." In: *Evolution; international journal of organic evolution* 59.1 (2005), pp. 24–37. ISSN: 0014-3820. URL: http://www.ncbi.nlm.nih.gov/pubmed/15792224.

[Ebe+07] Ingo Ebersberger et al. "Mapping Human Genetic Ancestry". In: *Molecular Biology and Evolution* 24.10 (2007), pp. 2266–2276. URL: http://dx.doi.org/10.1093/molbev/msm156.

[LY10] Liang Liu and Lili Yu. "Phybase: an R package for species tree analysis". In: *Bioinformatics* 26.7 (2010), pp. 962–963. URL: http://dx.doi.org/10.1093/bioinformatics/btq062.

[MDOMP16] Diego Mallo, Leonardo De Oliveira Martins, and David Posada. "SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees". In: *Systematic Biology* 65.2 (2016), pp. 334–344. URL: http://dx.doi.org/10.1093/sysbio/syv082.

[MM18] W. P. Maddison and D.R. Maddison. "Mesquite: a modular system for evolutionary analysis". In: *Version 3.51* (2018). URL: http://www.mesquiteproject.org/.

[PN88] P. Pamilo and M. Nei. "Relationships between gene trees and species trees." In: *Molecular Biology and Evolution* 5 (1988), 568583.

[Pol+06] Daniel A. Pollard et al. "Widespread discordance of gene trees with species tree in drosophila: Evidence for incomplete lineage sorting". In: *PLoS Genetics* 2.10 (2006), pp. 1634–1647. ISSN: 15537390.

[Ree+10] J. Reece et al. *Campbell Biology, 9th Edition*. Pearson, 2010.

[RY03] Bruce Rannala and Ziheng Yang. "Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci." In: *Genetics* 164 (2003), pp. 1645–1656.

[SLA16] Claudia Solís-Lemus and Cécile Ané. "Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting". In: *PLoS Genetics* 12.3 (2016). ISSN: 15537404.

[Syr+05] John Syring et al. "Evolutionary relationships among Pinus (Pinaceae) subsections inferred from multiple low-copy nuclear loci". In: *American Journal of Botany* 92.12 (2005), pp. 2086–2100. ISSN: 00029122.

[Wak08] John Wakeley. *Coalescent Theory: An Introduction*. Vol. 58. Roberts and Company Publishers, 2008, pp. 836–845. ISBN: 0974707759.

[Zhu+15] Sha Zhu et al. "Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees". In: *BMC Bioinformatics* 16.1 (2015), p. 292. ISSN: 1471-2105. URL: https://doi.org/10.1186/s12859-015-0721-y.

[R C13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.

7. Appendix

```
######### needs: ape, rSymPy, pryr #########
## Composition of the funtions to obtain the inverses of the population functions
inversi=function(x) 1/x
  invpopfun=list()
  for(i in  1:length(popfun))
        {
               invpopfun[[i]]= pryr::compose(inversi,popfun[[i]])
        }

##  This funtion returns the edges above the MRCA of taxon 1 and taxon 2  (note that the taxa can be the same)
edgesabove=function(t,tx1,tx2)
{
        MRCA=mrca(t, full = FALSE)
        MRCA=MRCA[tx1,tx2]
         ancestry=MRCA
         flag= length(which(t$edge[,2]==MRCA))
       p=1
       while(flag==1)
        {
               ancestry[p+1]=t$edge[which(t$edge[,2]==ancestry[p])]
```

```
                flag=length(t$edge[which(t$edge[,2]==ancestry[p+1])])
                 p=p+1
         }
         edge.ancestry=1
         for( i in  1:length(ancestry))
        {
                if(length(which(t$edge[,2]==ancestry[i]))==1)
                edge.ancestry[i] =  which(t$edge[,2]==ancestry[i])
                else{  edge.ancestry[i]=nrow(t$edge) +1}
        }
         return(edge.ancestry)
}

## This function gives the density of pairwise distances
pairwisedist= function(tx1,tx2,popfun,l,t)
{
  if(tx1==tx2){ warning('Taxon must be different')}
  edgs=t$edge.length
  ancestry=edgesabove(t,tx1,tx2)
  ea1=edgesabove(t,tx1,tx1)
  ea2=edgesabove(t,tx2,tx2)
  t1=ea1[!ea1 %in% ancestry]
  t2=ea2[!ea2 %in% ancestry]
  disjoint=sum(t$edge.length[t1],t$edge.length[t2])
  if(l<disjoint)
        {
           return(0)
        }
  nocoal=1
  if(length(ancestry)==1)
        {
          return( 1/(2*popfun[[ancestry[1]]](l))*exp(-1/2*integrate(invpopfun[[ancestry[1]]],0,l-disjoint)$value))
        }
 for(i in 1:(length(ancestry)-1))
        {
         if(l<disjoint+2*edgs[ancestry[i]])
           {
  return( 1/(2*popfun[[ancestry[i]]](l))*exp(-1/2*integrate(invpopfun[[ancestry[i]]],0,l-disjoint)$value)*nocoal)
      }
        disjoint=disjoint+2*edgs[ancestry[i]]
        nocoal=nocoal*exp(-integrate(invpopfun[[ancestry[i]]],0,edgs[ancestry[i]])$value)
        if(i==(length(ancestry)-1))
          {
  return( 1/(2*popfun[[ancestry[i+1]]](l))*exp(-1/2*integrate(invpopfun[[ancestry[i+1]]],0,l-disjoint)$value)*nocoal)
          }
   }
}
```